Developing Achievement Levels on the 2009 National Assessment of Educational Progress in Science for Grades Four, Eight, and Twelve

Process Report

Presented by ACT, Inc.
July 2010

Redacted by:
National Assessment Governing Board

National Assessment Governing Board
2009–2010

Committee on Standards, Design, and Methodology

Developing Achievement Levels on the 2009 National Assessment of Educational Progress in Science for Grades Four, Eight, and Twelve

Process Report

The work for this report was conducted by ACT, Inc., under contract ED-08-CO-0143 with the National Assessment Governing Board.

Process Report
Table of Contents

EXECUTIVE SUMMARY

DEVELOPING ACHIEVEMENT LEVELS FOR THE 2009 NAEP IN SCIENCE FOR GRADES 4, 8, AND 12: PROCESS REPORT

APPENDICES

List of Tables

List of Figures`

Page

EXECUTIVE SUMMARY

OVERVIEW

This report describes the process and outcomes of a meeting that was held in January 2010 to set achievement levels for the 2009 National Assessment of Educational Progress (NAEP) in science for grades 4, 8, and 12. The meeting was conducted by ACT, Inc., under contract with the National Assessment Governing Board. The contract calls for ACT to conduct Achievement Level Setting (ALS) activities consistent with Board policies and to develop recommendations for setting achievement levels. The actual setting of achievement levels is a policy judgment by the Governing Board, based on contractor recommendations. ACT bases its recommendations for achievement levels on evidence that the ALS process had procedural validity and was reliable. This report is a summary of such evidence.

In addition to describing the ALS meeting process, the report presents the recommended Achievement Level Descriptions (ALDs) and cut scores, and identifies items that may be used to illustrate what students in the achievement levels know and can do (referred to as exemplar items). Also described in the report are a Pilot Study that preceded the ALS meeting and activities to address three research questions posed by the Governing Board in their Statement of Work (NAGB, 2008a).

Additional information about the project may be found in the Technical Report (ACT, 2010), which documents technical advice ACT received in the project and data analysis procedures used throughout the process.

BACKGROUND

The National Assessment Governing Board has been setting achievement levels for grades and subject areas in the NAEP since 1992. Achievement levels have been set for the National Assessments in civics, economics, U.S. geography, U.S. history, mathematics, reading, science, and writing. As currently specified by Governing Board policy, there are two phases to the NAEP ALS process (National Assessment Governing Board, 2008b). In phase 1, grade-specific and subject-specific ALDs are developed from general policy definitions for three achievement levels—Basic, Proficient, and Advanced. The ALDs represent what students in the achievement levels should know and be able to do. In phase 2, the ALDs are translated into cut scores. Phase 1 occurs before the ALS meeting. Phase 2 is the ALS meeting.

Achievement levels have become the most publicly visible aspect of the NAEP, also known as The Nation’s Report Card. Achievement level percentages—the percent of students in each achievement level and the percent at or above each achievement level—show how students are performing relative to what students should know and be able to do. Trends in achievement level percentages have become a major resource to educators and policymakers assessing the nation’s progress toward its educational goals.

Achievement levels for science in grades 4, 8, and 12 were originally set for the 1996 Science NAEP. The framework for the 2009 Science NAEP, developed and approved by the Governing Board in 2005, is significantly different from the previous framework used for the 1996, 2000, and 2005 science assessments (NAGB, 2008c). The Governing

Board's policy is to update the achievement levels as needed, typically when assessment frameworks are updated. Because achievement scores on tests constructed from different frameworks are fundamentally not comparable, the Governing Board decided to discontinue the trend of comparable science results, and start a new trend using results from the 2009 NAEP in science for grades 4, 8, and 12. Consequently, it is inappropriate to compare achievement scores on the 2009 assessment to those from previous science assessments. Unfortunately, this will not prevent comparisons of the achievement level percentages from occurring.

In keeping with the Governing Board's procurement (NAGB, 2008a) which required the offeror to design and implement a bookmark-type methodology (Mitzel, Lewis, Patz, & Green, 2001) for setting achievement levels on the 2009 NAEP in science for grades 4, 8, and 12, ACT implemented the Mapmark with Whole Booklet Feedback standard-setting method (ACT, 2007). The Mapmark with Whole Booklet Feedback method is a three-round bookmark-based process with the provision of whole booklet feedback (i.e., actual student test booklets) in the second round. ACT's Mapmark method improves the Bookmark process with the use of item maps that illustrate the relative difficulty of all items within the assessment pool by locating each item on a score scale where a given Response Probability (RP) (in this case, 0.67) is met (Masters, Adams, & Lokan, 1994). The Mapmark with Whole Booklet Feedback procedure was the method used for setting achievement levels on the 2006 NAEP for grade 12 economics.

A Pilot Study was conducted in October 2009, prior to the ALS meeting. The purpose of the Pilot Study was to implement the standard-setting process planned for the ALS meeting and to gather information on areas for improvement. The Pilot Study and the ALS meeting used results from the 2009 science assessment. Both studies used the same process and content facilitators for each grade level. In addition, the studies used the same techniques for obtaining panelists, developing briefing materials, and orienting panelists to the NAEP and the general purposes of standard setting. These processes, while part of a standard-setting design, are not specific to a particular ALS procedure. ACT presented the results of the Pilot Study and its recommendations for process improvement to its Technical Advisory Committee on Standard Setting (TACSS) on November 5–6, 2009, and to the Governing Board's Committee on Standards, Design, and Methodology (COSDAM) on November 20, 2010.

As recommended by the Governing Board, ACT implemented the Mapmark with Whole Booklet Feedback method in the operational ALS meeting.

ALS Meeting

The ALS meeting lasted three and a half days, January 28–31, 2010 (Thursday through Sunday). It was conducted at the Westin Riverwalk Hotel in San Antonio, Texas. Sessions generally started at 8:00 a.m. and lasted until 5:00 p.m. or 6:00 p.m., except the last day, which adjourned at noon. The ALS agenda is in Appendix A.

The Panelists

Policies of the Governing Board regarding qualifications of panel members and composition of panels were followed. Panelists were selected from a group of nominees. These nominations were submitted by nominators who had been chosen to provide a representative sample of the educational community. Nominations came from a variety of sources, including schools, districts, state education agencies, colleges, chambers of commerce, science organizations and museums, and businesses. The businesses contacted were those companies that would be likely to have a large number of employees with science-related backgrounds, including drug companies, high-tech manufacturers, and engineering firms. At the school level, a stratified random sample of 1,000 schools was selected at each of the three grade levels: 4, 8, and 12. Additionally, a random sample of 6,000 teachers was drawn from a national database of science teachers. In all, over 12,000 solicitations for nominations were sent. Each nominator was able to nominate for any of the three panelist types—teachers, nonteacher educators, and general public—although in certain instances, specific panelist types were requested (e.g., nonteacher educators from colleges and general public from businesses). A total of 674 nominations were received: 165 at grade 4, 223 at grade 8, and 286 at grade 12. From this pool, panelists were selected for the Pilot Study and the ALS meeting.

Panels for the ALS meeting were recruited to meet Governing Board specifications that (1) the panels be broadly representative by gender, race/ethnicity, and region of the country, (2) that two-thirds (70%) be educators and one-third (30%) noneducators, with classroom teachers comprising 55% of the educator group, (3) that all panelists have science training to qualify. Recruited teachers had at least 5 years of experience; were active in local, state, and national science organizations; had been nominated for or won teaching awards at the local, state, and national level; used innovative teaching techniques; and were held in high regard by their peers. Recruited nonteacher educators had degrees in science; were involved with local, state, and national science organizations; served as a curriculum specialist at the local, state, or national level; and were former teachers for the appropriate grade level. General public panelists had degrees in science and were currently working in a science-related field; had children either at or recently at the appropriate grade level; volunteered at their school; were a civic or public leader; or had been recognized for civic work. Table 1 shows the profile of the 85 panelists who participated in the ALS meeting.

Table 1: ALS panels by panelist type, gender, race/ethnicity, and geographic region

Panel Demographics	Number (%) of Panelists		
	Grade 4	**Grade 8**	**Grade 12**
Panelist Type			
Teacher	16 (53)	13 (48)	17 (61)
Nonteacher Educator	5 (17)	6 (22)	5 (18)
General Public	9 (30)	8 (30)	6 (21)
Gender			
Female	18 (60)	14 (52)	11 (39)
Male	12 (40)	13 (48)	17 (61)
Race/Ethnicity			
Minority	4 (13)	4 (15)	7 (25)
Nonminority	26 (87)	23 (85)	21 (75)
Geographic Region			
Midwest	7 (23)	6 (22)	7 (25)
Northeast	5 (17)	7 (26)	5 (18)
South	11 (37)	8 (30)	9 (32)
West	7 (23)	6 (22)	7 (25)
Total	**30**	**27**	**28**

Design Factors

Rater groups and table groups were design factors in the ALS process. For each grade level, panelists were divided into two rater groups (Group A and Group B) and six table groups. Each rater group was divided into three table groups of five or six panelists each. The demographic attributes and content expertise of panelists were considered when assigning members to rater groups and tables; otherwise the assignments were random. The goal was to have rater groups as equal as possible with respect to panelist type, gender, race/ethnicity, and geographic region, and to have at least two panelists with expertise in each of the three content areas at each table. Group A and Group B worked with different but equivalent and overlapping item-rating pools. For each grade level, the two item-rating pools each contained 60–61% of the items in the 2009 assessment. Combined, they represented 100% of the items. Separate item rating pools were used to reduce the number of items any one panelist needed to consider.

The ALS Meeting Process

The Mapmark with Whole Booklet Feedback method was implemented in the operational ALS meeting for science in grades 4, 8, and 12. The method uses a bookmark procedure (Mitzel et. al., 2001) with the addition of item maps in all three rounds, whole booklet feedback in round 2, and consequences data in round 3. The method, which was implemented in the same manner for all three grades, is described generally in this executive summary and in detail in the pages that follow.

Orientation

The ALS meeting began with a motivational speech by Governing Board member Senator van de Putte from Texas. Following her welcome, the project director introduced the ACT staff, the process and content facilitators, and the observers for the meeting and described the process for selecting panelists. Next, Susan Loomis, the Governing Board's contracting officer's representative (COR), reviewed the history, organizational structure, and key policies of the NAEP as well as the purpose of setting achievement levels. The senator also presented the 2005 science achievement level percentages. Following this orientation, the panelists took a form of the 2009 NAEP assessment appropriate for their grade group and scored their own performance. They were then instructed on the Science Framework and introduced to the Mapmark with Whole Booklet Feedback method. The primary steps in the ALS process were described and training was provided for some of the key materials and concepts.

Round 1

The first round of the ALS included a review of the items with identification of what a student needs to know and be able to do to get an item correct; a presentation of the ALDs; and a demonstration of how to place bookmarks for each achievement level. Bookmark placements were done one achievement level at a time starting with Proficient, then Basic, then Advanced.

Round 1 began with an exercise aimed at familiarizing panelists with the items in the assessment and, subsequently, the gradient of difficulty in the assessment content. In this exercise, panelists reviewed the items in their item-rating pool in a Constructed Response Ordered Item Book (CROIB), which contained the constructed-response (CR) items, and in an Ordered Item Book (OIB), which presented the items (both multiple-choice (MC) and CR) in order of difficulty from easiest to hardest. More specifically, items were ordered in the OIB by the scale value or estimate of student ability that corresponds to a 0.67 probability of correctly answering the item or getting full credit on the item based on the results from the 2009 administration of the assessment. Items were also presented on the item maps by the scale values that correspond to a 0.67 probability of getting credit on the item when scoring at a given score level (see Appendix B). The RP value of 0.67 is based on an item response theory (IRT) model.

Each stage of the item review was designed to help panelists gain a clearer sense and a common understanding of what the assessment measures. The item review started with a grade-group review of four of the common CR items in the CROIB. In this first stage, which took approximately 1 hour, the facilitators used the four CR items to model the item review process for the panelists. In the next stage, the panelists were given approximately 4 hours to review the remaining CR items in their item pool within their table group. Following review of the CR items, panelists had approximately 2 hours for independent review of 80% of the MC items in their item pool. The item review task ended with a 2 hour table-group review of all the items, CR and MC, in the group's item pool. During the item review task, panelists were asked to identify for each item what students need to know and be able to do in order to correctly answer an item or to reach a given score point.

Following the item review, the content facilitators then led grade-group discussions of the ALDs. After panelists came to an understanding of what characterizes Basic, Proficient, and Advanced student performance, they were instructed in how to place their bookmark to represent their understanding of what students need to know and be able to do in order to be just qualified to be in an achievement level. Mastery was defined as a 0.67 probability of a correct response on the item by the minimally-qualified or borderline student. Panelists were told that students would be expected to have about a 67%

chance of correctly answering the items at the scale value where the bookmark was placed in the OIB, a higher likelihood of correctly answering items below that scale value, and a lesser likelihood of correctly answering items above that scale value.

Round 2

In round 2, panelists were provided with the three median cut scores (hereafter called *group cut scores*) from round 1, the distribution of all individual panelists' cut scores, and with actual student test booklets (i.e., whole booklet feedback) to help with their understanding of what students in the levels that they defined in round 1 can do. Three test forms were used for each grade—one test form for Group A, one for Group B, and one common to both groups. Each table within a grade group received 20 booklets: 10 for the common form and 10 for the group-specific form. Two booklets of each form represented performance at each group cut score and one booklet of each form represented performance at the middle of each achievement level range, including Below Basic. Through independent review and table discussion of the student booklets, panelists evaluated student performance with respect to these points and in relation to their own cut scores for each level. Following this review, panelists selected a scale value representing their cut score for each level. They started with either the student booklet information or the OIB, whichever they felt more comfortable with, and then tried different cut scores until they were satisfied with the meaning of the cut score in both frames of reference.

Round 3

In round 3, panelists were provided with the group cut scores from round 2, the distribution of all individual panelists' cut scores, and the proportion of students performing at or above and within each achievement level based on the round 2 group cut scores. Panelists discussed these *consequences data*. They were asked to consider the consequences data as a reality check on their cut score selection and were given an opportunity to adjust their cut scores if appropriate.

Post-Round Activities

Following round 3, panelists were again given feedback from the previous round in the form of the final group cut scores, the cut score distribution, and consequences data based on those cut scores. On a consequences questionnaire, they indicated their reactions to the consequences data, including whether they wished to recommend alternative cut scores or would like to leave the cut scores unchanged. Finally, they provided recommendations concerning the selection of exemplar items for the achievement levels. Panelists were instructed to discuss each potential exemplar item with their table group, but they were to provide independent ratings on the basis of whether the science content and practices required by the item seemed appropriately matched to the achievement level. They were instructed to consult their ALDs in this task and to rate each item *Very Good*, *OK*, or *Do Not Use* as an exemplar.

Evaluations of the ALS Meeting Process

Procedural validity of the ALS process was evaluated through process evaluation questionnaires given to panelists at the conclusion of each round and each day. Many of the questions had been used in the Pilot Study and in previous ALS meetings. A detailed summary of responses is contained in the full report. However, the data in Table 2 are representative of the fact that the Mapmark with Whole Booklet Feedback process was well implemented. Average responses, all on a 1–5 Likert scale, from the ALS meeting and the Pilot Study are shown for comparison.

Table 2: Average responses to process evaluation questions

Question	Meeting	Average ALS Panelist Response		
		Grade 4	Grade 8	Grade 12
The instructions on what I was to do during each round were . . . (5 = *absolutely clear*)	ALS	4.50	4.19	4.64
	Pilot	3.95	4.37	4.21
My understanding of the tasks I was to accomplish during each round was . . . (5 = *totally adequate*)	ALS	4.63	4.46	4.79
	Pilot	4.19	4.37	4.53
The most accurate description of my level of confidence in the cut score recommendations I provided was . . . (5 = *totally confident*)	ALS	4.63	4.56	4.78
	Pilot	4.48	4.53	4.63
The amount of time I had to complete the tasks I was to accomplish during each round was . . . (3 = *about right*)	ALS	2.90	3.04	3.21
	Pilot	2.76	2.84	2.58
I would describe the effectiveness of the Achievement Level Setting method as . . . (5 = *highly effective*)	ALS	4.38	4.15	4.64
	Pilot	3.95	4.32	4.05
I feel this ALS process provided me an opportunity to use my best judgment to recommend cut scores for the NAEP science assessment . . . (5 = *to a great extent*)	ALS	4.70	4.89	4.82
	Pilot	4.62	4.84	4.68

On the key process evaluation questions in Table 2, where 1 is least favorable and 5 is most favorable, the average response has historically been 4.0 or higher. This was the case with both the grade 12 economics ALS process and the grade 12 mathematics ALS process. On ratings of clarity of instructions, understanding of their tasks, confidence in cut scores, effectiveness of ALS method, and opportunity to use their best judgment, the science ALS process performed well in relation to previous ALS processes.

The ALS process was also evaluated on the basis of the following criteria:

- reasonable variability of cut scores across panelists and rounds,
- absence of extreme reactions to consequences data (the percent of students at or above each achievement level), and
- adequate number of exemplar items for each achievement level.

Evaluations of the ALS process on these criteria were positive. Details are provided in the full report and in other sections of this executive summary.

ALS Process Outcomes

The ALS process consists of all activities leading up to the setting of achievement levels by the Governing Board. In setting the achievement levels, the Governing Board adopts three major outcomes of the ALS process: ALDs, cut scores, and exemplar items. Exemplar items are used to illustrate what students in each achievement level know and can do.

Achievement Level Descriptions

The development of ALDs in this project conformed to the two-phase process described earlier in that the ALDs were developed before the ALS meeting and their development was outside the scope of this project. The ALDs were developed by science experts working with the Governing Board and were provided to ACT for use in this standard-setting project. They were used in both the Pilot Study and at the ALS meeting. The ALDs for grades 4, 8, and 12 are reproduced in Appendix C.

When ACT reported on the ALS meeting to the Governing Board at their March 2010 meeting, ACT endorsed the ALDs used in the ALS meeting. As reported in Table 3, the ALS panelists reported that they understood the ALDs and found them useful for setting the cut scores. Based on this and on the fact that the ALDs served as the basis from which the cut scores were established, ACT did not recommend any changes or modifications to the ALDs.

Table 3: Average responses to ALD process evaluation questions

		Mean ALS Panelist Response		
Question	**Round**	**Grade 4**	**Grade 8**	**Grade 12**
At the time I provided the round 1 bookmark placements, my understanding of the Basic ALD was . . . (5 = *totally adequate*)	1	4.03	4.11	4.41
At the time I provided the round 1 bookmark placements, my understanding of the Proficient ALD was . . . (5 = *totally adequate*)	1	4.07	4.26	4.37
At the time I provided the round 1 bookmark placements, my understanding of the Advanced ALD was . . . (5 = totally adequate)	1	4.03	4.19	4.33
At the time I placed my round __ cut score recommendations, my understanding of the ALDs was . . . (5 = *totally adequate*)	2	4.66	4.59	4.63
	3	4.80	4.70	4.89
During the ALS process, I found the ALDs . . . (5 = *very helpful*)	Post-round	4.87	4.93	4.75

Cut Scores

ACT recommended that the Governing Board adopt the group cut scores from round 3 of the ALS meeting. (See Table 4. Cut scores are reported on a converted NAEP scale called the *ACT NAEP-like scale*). This recommendation was based on the conclusion of ACT and ACT's TACSS that the ALS process had procedural validity and produced reliable results across panelist type and group. Also, round 3 cut scores are based on all of the information that ACT recommends be considered by panelists in adopting cut scores, including student performance data. In addition, the differences in cut scores obtained from the ALS meeting and the Pilot Study were generally small. The analysis of procedural and internal consistency data from the ALS meeting suggest that the panelists were well qualified and the method was conducted well, understood by panelists, and not unduly impacted by variation in the panelists. The conclusion was that the results of the ALS meeting are reasonable. The percentages of students in each of the achievement level groups using the ALS group cut scores are in Table 5.

Table 4: ALS cut scores* by grade and achievement level

Grade	Basic	Proficient	Advanced
4	328	376	447
8	570	598	647
12	785	820	866

*The ACT NAEP-like scales have means (SDs) of 364 (33), 579 (33), and 793 (33) for grades 4, 8, and 12, respectively.

Table 5: ALS percentages by grade and achievement level

Grade	Achievement Level	Percentage
4	At or Above Advanced	0.1
	At or Above Proficient	39.5
	At or Above Basic	85.9
	Below Basic	14.1
8	At or Above Advanced	0.7
	At or Above Proficient	30.3
	At or Above Basic	63.5
	Below Basic	36.5
12	At or Above Advanced	0.8
	At or Above Proficient	20.9
	At or Above Basic	60.2
	Below Basic	39.8

Exemplar Items

Following round 3 of the ALS meeting, panelists provided input on the suitability of selected items for illustrating what students in the achievement levels, as defined by the round 3 cut scores, know and can do. Two blocks of items had been designated for possible public release and panelists classified the MC items and CR score points in those two blocks that mapped within each achievement level based on the final group cut scores. The statistical criteria ACT used to associate items with achievement levels for the rating task used the RP criterion panelists had used to place their bookmarks and determine cut scores. All potential exemplars associated with scale values within an achievement level using this criterion were selected for review by the panelists. Panelists individually rated the items as *Very Good*, *OK*, or *Do Not Use* as an exemplar based on the match between item content and the ALDs.

The results of the exemplar item rating task for each achievement level for each grade are given in Appendix D and summarized in Table 6. ACT suggested that the Governing

Board use items rated by 50% or more of the panelists as *Very Good* and by fewer than 30% of the panelists as *Do Not Use* as exemplars in the reporting of NAEP results.

Table 6: Number of potential exemplar items rated by 50% or more as Very Good and by fewer than 30% as Do Not Use*

	Achievement Level		
Grade	**Basic**	**Proficient**	**Advanced**
4	4 (12)	9 (24)	4 (5)
8	7 (11)	6 (23)	5 (12)
12	3 (10)	9 (22)	7 (11)

*Total number of score points mapping within the achievement level is given in ().

RECOMMENDATIONS

ACT's principal recommendations concern the three outcomes of the ALS process—ALDs, cut scores, and exemplar items.

- ACT endorses the ALDs.
- ACT recommends the cut scores from round 3 of the ALS meeting.
- ACT recommends that the Governing Board use the lists of items and panelists' ratings from the ALS meeting, coupled with other information, in the process of selecting exemplar items.

The basis for these recommendations is provided in the full report.

DEVELOPING ACHIEVEMENT LEVELS FOR THE 2009 NAEP IN SCIENCE FOR GRADES 4, 8, AND 12: PROCESS REPORT

INTRODUCTION

Background on NAEP Achievement Level Setting Activities

Achievement levels on the National Assessment of Educational Progress (NAEP) are intended to help teachers, parents, educators, policymakers, and the general public understand how students in the United States are performing on the NAEP relative to what students should know and be able to do. Public Law 100-279 mandates the National Assessment Governing Board to identify "appropriate achievement goals for each grade or age in each subject area to be tested" under the National Assessment. Governing Board policy (NAGB, 2008b) specifies three achievement levels—Basic, Proficient, and Advanced—and states that the purpose of these levels is to make NAEP data more understandable to the general user, parents, policymakers, and educators alike. Achievement levels have been set for NAEP assessments in civics, economics, U.S. geography, U.S. history, mathematics, reading, science, and writing. Achievement level percentages—the percent of students at or above each achievement level—have become the principal means by which educational policymakers assess the nation's progress in meeting its educational goals.

There are three components of NAEP achievement levels: ALDs, cut scores, and exemplar items. ALDs are descriptions specific to the subjects and grades assessed by NAEP (4th, 8th, and 12th) of what students should know and be able to do in each level—Basic, Proficient, and Advanced. Cut scores are numerical representations of the lower borderline of each level. Exemplar items are matched with achievement levels in order to illustrate the kinds of knowledge and skills required for performance within each level.

As currently specified by Governing Board policy (NAGB, 2008b), there are two phases to the NAEP Achievement Level Setting (ALS) process. In phase 1, grade- and subject-specific ALDs are developed from general policy definitions. In phase 2, the ALDs are translated into cut scores and exemplar items to represent the achievement levels are identified. Phase 2 has traditionally been performed in an ALS meeting by a panel of teachers, nonteacher educators, and representatives of the general public. The targeted percentages of these types of panelists are, respectively, 55%, 15%, and 30%. This is in keeping with Governing Board policy that the development of achievement levels shall be a widely inclusive activity. The Governing Board may call for field trials, pilot studies, and other research activities designed to improve the standard-setting process and the way that standard-setting results are reported.

Ultimately, the setting of achievement levels is an exercise of policy judgment by the Governing Board. Key criteria in the Governing Board's policy judgment are the validity and reliability of the ALS process and the apparent reasonableness of the results. The Governing Board specifies that the final reports for ALS activities are to serve as the principal means of documenting these criteria for specialists in the field as well as for the general public.

Background on the Current Project

The development of the Science Framework, the assessments, and ALDs, and consequent administration of the assessments to national samples of 4th, 8th, and 12th grade students in spring 2009 led the Governing Board to issue a procurement for establishing cut scores and identifying exemplar items for achievement levels in science for grades 4, 8, and 12.

Achievement levels for science in grades 4, 8, and 12 were originally set for the 1996 Science NAEP. The framework for the 2009 Science NAEP (NAGB, 2008c), developed and approved by the Governing Board in 2005, is significantly different from the previous framework used for the 1996, 2000, and 2005 science assessments. NAGB policy is to update the achievement levels as needed, typically when assessment frameworks are updated. Item statistics and student distribution data for all ALS activities in this project are based on the results from the 2009 administration.

This report provides a detailed description of the method and outcomes of a meeting that was held in January 2010 to set achievement levels for the 2009 NAEP in science for grades 4, 8, and 12. It also describes a Pilot Study that was held in October 2009 as a try-out of the procedures that had been designed for the ALS meeting.

ACT proposed and implemented the Mapmark with Whole Booklet Feedback method for the science ALS meeting. The Mapmark method was developed by ACT for the grade 12 mathematics ALS project (ACT, 2005). That first implementation of the Mapmark method used holistic feedback in the form of domain-score feedback. Subsequently, ACT implemented Mapmark with holistic feedback in the form of whole booklet feedback for the grade 12 economics ALS (ACT, 2007). Mapmark is based on the Bookmark method which was introduced in 1996 (Lewis, Mitzel, & Green, 1996). Since then, Bookmark has become the most widely used standard-setting method in state assessments. ACT believes that the Bookmark method contains some very attractive features for setting standards, but that it can be improved with the use of spatially-representative item maps (Masters, Adams, & Lokan, 1994) and holistic feedback, such as whole booklet feedback (Loomis & Hanick, 2000). The Mapmark method used in this contract uses the Bookmark procedure (described by Mitzel, Lewis, Patz, & Green, 2001) in round 1, and provides holistic feedback in the form of student test booklets in round 2 and in the form of consequences data in round 3. In addition to the book of items ordered by difficulty used in Bookmark procedures, item maps are used in every round of Mapmark. An item map shows the test items arranged on a linear continuum representing both item difficulty and student achievement on the score scale (Appendix B).

ACT consulted with its Technical Advisory Committee on Standard Setting (TACSS) in all aspects of the project. The TACSS is a six-member group that collectively represents expertise in standard setting, science education, and experience with the NAEP. (See Appendix E for a list of the TACSS members.) The TACSS convened six times over the course of the project and provided input on key components of the project including the design of the ALS method, implementation of the method for the Pilot Study, results of the Pilot Study, implementation of the method for the ALS meeting based on results from the Pilot Study, results from the ALS meeting, and the formulation of conclusions and recommendations presented to the Governing Board's Committee on Standards, Design, and Methodology (COSDAM).

Contract Activities Prior to the ALS Meeting

Contract activities prior to the ALS meeting fall into two categories: (a) collection of public comment on the design document and (b) the Pilot Study. These activities are described in the following sections.

Public Comment

A website was established early during the design phase of the project in an attempt to obtain public comment on the Design Document. The website included the materials for comment and a survey that visitors could either complete online or download and mail to ACT after completing it. Feedback was solicited from professional associations, organizations, and stakeholder groups that will have a particular interest in the science ALS results. The website was advertised to members of these groups through listservs, newsletter announcements, and other communication channels.

Unfortunately, only one response was received. The respondent indicated that the criteria for grades 4, 8 and 12 were clearly stated and consistent with professional standards. Other comments obtained from this respondent addressed the layout and readability of the Design Document itself.

Pilot Study

The Pilot Study for the 2009 science ALS process was planned as a “dry run” for the operational ALS to determine whether modifications to training, instructions, materials, timing, and so forth were needed. The Pilot Study was conducted on October 8–11, 2009, at the Westin Riverwalk Hotel in San Antonio, Texas. The Mapmark with Whole Booklet Feedback standard-setting method was implemented. The method was essentially the same as the method ACT used to set achievement levels for the 2007 NAEP in grade 12 economics. Throughout the Pilot Study, ACT collected information about the reactions of panelists to the ALS process; and, following the Pilot Study, ACT collected feedback from the process facilitators, content facilitators, observers, and ACT staff. All this information was shared with TACSS and lead to adjustments in the process to assure smooth implementation of the methodology when used for the operational ALS meeting.

The essential elements of the Mapmark with Whole Booklet Feedback procedure are described in the ALS process section of this report. Based on lessons learned from the Pilot Study, three main changes were made to the process for the ALS meeting. First, all whole-group sessions were eliminated on days 2, 3, and 4 of the ALS meeting. The intent of these sessions was to help standardize the instructions for the panelists in all grades. However, there was a lot of redundancy between the whole-group sessions and the subsequent grade-group sessions. Thus, it was recommended that the whole-group sessions be dropped in order to gain more time for critical tasks such as item review. Second, the ALS agenda was revised to provide more time for the CR item review and the ALDs presentation sessions that were found to be rushed in the Pilot Study. Third, the CROIB and the OIB were modified to increase ease of use by panelists during item review. Those elements in the Pilot Study that differed from the ALS are described in Appendix F.

The Achievement Level Setting Process

ALS in NAEP refers to the overall process through which cut scores and exemplar items are obtained. The ALS meeting is just one part of the process. Activities leading up to the ALS meeting include the recruitment of panelists and mailing of advance materials.

Panelist Selection

ACT implemented the same basic design for selecting panelists to set achievement levels that ACT had used for the 2007 ALS process. This design was used for both the Pilot Study and the ALS meeting. Primary requirements based on NAGB policy were that the panel be broadly representative, and that 70% be educators and 30% non-educators. Moreover, classroom teachers should comprise 55% of the group. In addition to these primary requirements, both demographic characteristics and group size were key considerations in the selection of panelists.

In order to get a broad spectrum of panelists, several approaches were used to get nominations. Nominations of panelists were requested from a sample of school districts, teachers, state education associations, colleges/universities, and businesses and professional associations. Panelists were selected for recruitment from the sample of nominees. The sample of nominees was used for both the Pilot Study and ALS, and the same methods of selection were used for both. The following summary highlights the main features of each step in the process of selecting panelists to set achievement levels.

Selection of School Districts

Schools served as one of the basic units of sampling. A sample of schools was drawn for each grade to identify nominators of teachers, nonteacher educators, and the general public. The stratified random sample was drawn from the Market Data Retrieval (MDR) database of schools. The samples provided nominators for both the Pilot Study and the ALS meeting. The school samples were approximately proportional to the regional share of districts. Table 7 gives the regional proportions:

Table 7: Percentage of schools sampled by geographic region

	Geographic Region			
Grade	**Northeast**	**South**	**Midwest**	**West**
4	16	34	25	25
8	17	34	27	22
12	16	37	26	22

A total of 1,000 schools were sampled at each grade level, stratified by region and size of school. Table 8 gives the distribution of public and private schools sampled. Note that grade 8 has a higher percentage of private schools. This is a function of the number of schools of each type at each grade level. Private schools tend to include grades K–8. In public schools, grade 4 is typically located in a school that includes grades K–5 or K–6, with these elementary schools merging into larger entities at grades 6–8 or 7–9.

Table 8: Schools sampled by school type

School Type	Grade 4	Grade 8	Grade 12
Public	735	654	785
Private	265	346	215
Total	1,000	1,000	1,000

In addition to contacting schools, ACT also contacted 500 randomly selected school districts. The solicitation was addressed to the superintendent, and nominations for all three grades and all three panelist types were solicited. All 50 state departments of education were contacted. A random sample of 250 colleges was contacted specifically asking for nominations of college faculty in science education or teachers of first-year science classes (these would be considered nonteacher educators). In order to get nominations for the general public category, a list of science-related businesses were sent a mailing asking for nominations from their employees. Mailings were also sent to science and science education organizations and publications, and science centers and museums. Chambers of Commerce were also contacted. Finally, a random selection of almost 6,000 teachers, 2,000 each at grades 4, 8 and 12, was purchased from Market Data Retrieval's database of teachers, and each of these was sent information on nominations. A total of 12,617 individuals were contacted and asked to serve as nominators. Table 9 shows the distribution of nominators by type.

Table 9: Nominators contacted by type

Nominator Type	Number
Businesses (science related)	72
Chambers of Commerce	1,836
Colleges/Universities	250
School District Superintendents	1,000
School Principals	3,000
Science Centers and Museums	344
Science/Education Organizations & Publications	60
State Assessment Directors	52
State Science Supervisors	55
Science Teachers	5,948
Total	12,617

Nominators could submit candidates whom they judged to be well qualified to serve as standard-setting panelists. To submit a candidate, nominators had to fill out a questionnaire describing the candidate's qualifications (e.g., years of experience, professional honors and awards, degrees earned). They were encouraged to nominate members of minority groups. All nominators were permitted to nominate any type of panelist.

Selection of Panelists

Nominees represented a specific role (teacher, nonteacher educator, or member of the general public). A single pool of nominees was used for both the Pilot Study and the ALS. The Pilot Study sample was drawn from the nominees available at the time of sampling. ACT continued to accept nominations throughout the Pilot Study phase. Individuals that were contacted to participate in the Pilot Study that were unable to attend were returned to the nominee pool for possible selection for the ALS meeting. A total of 674 candidates were nominated to serve as potential panelists.

A computerized algorithm was developed to select panelists from the pool of nominees. Nominees were evaluated according to their qualifications based on information provided on the nomination form (e.g., years of experience, professional honors and awards, degrees earned). The selection program was designed to yield panels with:

- 55% of the members representing grade level classroom teachers,
- 15% of the members representing nonteacher educators,
- 30% of the members representing the general public,
- 20% of the members from diverse minority racial/ethnic groups,
- up to 50% of the members male, and
- appropriate percentage (based on census population) of the members representing each of the four NAEP regions.

Ninety panelists were required for the ALS panels, 30 per grade level. One hundred and ten persons were selected from the nominee pool and contacted about serving as an ALS panelist. Some of the persons who were selected declined or did not respond, and two became incapacitated during the meeting (see Table 10). Consequently, a total of 85 panelists participated in the ALS (see Tables 11–15). A list of the panelists who participated in the ALS is in Appendix H.

Table 10: Participation of candidates

Candidate	Number	Percent of Invited
Received Invitation	110	100
Declined upon invitation	6	5
Did not respond	8	7
Declined after agreeing	9	8
Finally agreed	87	79
Became ill at the meeting	2	2
Participated	85	77

Table 11: Grade 4 ALS panel by panelist type, gender, and race/ethnicity

	Male		Female		
Panelist Type	**Minority**	**Nonminority**	**Minority**	**Nonminority**	**Total (%)**
Teacher	0	4	2	10	16 (53)
Nonteacher Educator	0	2	2	1	5 (17)
General Public	0	6	0	3	9 (30)
Total	**0**	**12**	**4**	**14**	**30**
Percent	**40**		**60**		

Table 12: Grade 8 ALS panel by panelist type, gender, and race/ethnicity

Panelist Type	Male		Female		
	Minority	Nonminority	Minority	Nonminority	Total (%)
Teacher	0	4	1	8	13 (48)
Nonteacher Educator	1	2	1	2	6 (22)
General Public	0	6	1	1	8 (30)
Total	**1**	**12**	**3**	**11**	**27**
Percent	**48**		**52**		

Table 13: Grade 12 ALS panel by panelist type, gender, and race/ethnicity

Panelist Type	Male		Female		
	Minority	Nonminority	Minority	Nonminority	Total (%)
Teacher	4	5	2	6	17 (61)
Nonteacher Educator	0	3	1	1	5 (18)
General Public	0	5	0	1	6 (21)
Total	**4**	**13**	**3**	**8**	**28**
Percent	**61**		**39**		

Table 14: ALS panels by region

Geographic Region	Grade 4	Grade 8	Grade 12	Total
Midwest	7 (23)	6 (22)	7 (25)	15
Northeast	5 (17)	7 (26)	5 (18)	12
South	11 (37)	8 (30)	9 (32)	23
West	7 (23)	6 (22)	7 (25)	10
Total	**30**	**27**	**28**	**85**

Table 15: ALS panels by content area expertise

Content Area	Grade 4		Grade 8		Grade 12		
	Teachers	Other	Teachers	Other	Teachers	Other	Total
Earth & Space Sciences	14	10	9	4	6	3	46
Life Science	15	11	9	10	11	5	61
Physical Science	16	7	8	10	12	8	61
Total	**16**	**14**	**13**	**14**	**17**	**11**	

Advance Materials

Before the ALS meeting, all panelists were mailed materials that contained important background information on setting achievement levels. On November 17th, panelists were sent a letter thanking them for agreeing to participate in the ALS meeting. The letter also briefly described the purpose of the meeting and provided information regarding the location of the meeting and expense detail. Panelists were sent another letter on December 10th providing travel arrangement information. On January 8th, panelists were sent a packet of advance materials. The cover letter briefly described the enclosed materials, specified sections of the framework to read in advance of the meeting, and provided detailed hotel information, guidance on appropriate dress for the meeting, transportation information, and expense reimbursement information. Enclosures were:

- hotel information including directions,
- ALS Meeting Agenda,
- Briefing Booklet,
- NAEP 2009 Science Framework,
- NAEP 2009 Science ALDs,
- Confidentiality agreement,
- request for reimbursement form, and
- local news release form

A form of the briefing booklet was first used by ACT for the 1994 ALS process. The briefing booklet included a description of the goals and objectives for the process and a brief description of each step in the process; and it defined key terms used in the standard setting meeting. The briefing booklet used for the science Pilot Study was modified for the ALS meeting to reflect the changes made to the agenda and process based on feedback from the Pilot Study. A copy of the ALS Briefing Booklet is provided in Appendix I.

The ALS Meeting

The ALS meeting lasted three and a half days, January 28–31 (Thursday through Sunday). It was conducted at the Westin Riverwalk Hotel in San Antonio, Texas. Sessions generally started at 8:00 a.m. and lasted until 5:30 p.m. or 6:00 p.m., except the last day, which adjourned at 12:15 p.m. The agenda is shown in Appendix A.

Design Factors

Prior to the meeting, panelists were assigned to two groups of about 15 persons each: Group A and Group B. Each group rated a different, but overlapping, set of items as explained in the next section. Each group was further divided into three table groups of four or five panelists each. The demographic attributes and content expertise of panelists were considered when assigning members to groups and tables; otherwise the assignments were random. The groups were divided to be as equivalent as possible. The goal was to have groups and tables as equal as possible with respect to panelist type, gender, race/ethnicity, and geographic region and to have at least 2 panelist with expertise in each of the three content areas at each table.

Group A and Group B worked with different but equivalent and overlapping item-rating pools. For each grade level, the two item-rating pools each contained 60–61% of the items in the 2009 assessment. Combined, they represented 100%. Separate item rating pools were used to reduce the number of items any one panelist needed to consider.

Item Pool Division

All operational items (test questions) in the 2009 science assessment pools for grades 4, 8, and 12 were used in the ALS meeting. For grade 4, there were 141 items representing 196 score points; for grade 8, there were 162 items representing 249 score points; and for grade 12, there were 179 items representing 245 score points. Items were in two basic formats: multiple choice and constructed response.

The item pool was divided into equivalent, but overlapping, pools for rater groups A and B. Equivalence was evaluated with regard to: (a) mean and variation of item difficulty, (b) percent of items/score points of each type, (c) representation of science content areas, (d) representation of CR items by number of score points, and (e) representation of science practices.

The equivalence criteria were met, to the extent possible, by assigning intact blocks of items to the two item-rating pools. Blocks are sets of approximately 16 items created for purposes of test form construction to require approximately 25 minutes of student response time. Each student test booklet contained two blocks of cognitive assessment items and some background information questions. The 2009 assessment consisted of nine, ten, and eleven blocks for grades 4, 8, and 12 respectively. The two item-rating pools for each grade had two blocks in common. The common blocks were ones being considered for release to the public after the assessment. For grade 4, five intact blocks were assigned to each item-rating pool and the items in the remaining block were divided between the two pools. For grade 8, six intact blocks were assigned to each item-rating pool. For grade 12, six intact blocks were assigned to each item-rating pool and the items in the remaining block were divided between the two pools. The item pool division used in the ALS was identical to that used for the Pilot Study. The two item-rating pools for each grade are summarized in Appendix J with regard to the key characteristics listed above. Also shown in Appendix J is the number of unique, common, and total items assigned to each rater group for each grade.

Facilitation, Room Setup, and Observers

Christina Hamme Peterson served as the lead process facilitator and as the process facilitator for grade 12. Rosanne Cook was the process facilitator for grade 8, and Nancy Allen was the process facilitator for grade 4. All three facilitators had served as process facilitators for the Pilot Study and were experienced in the procedures performed in the ALS meeting. Peterson was the project director and a process facilitator for the 2006 NAEP grade 12 economics ALS. She has taught, set standards for grade 12 Biology for the Department of Defense Education Activity (DoDEA), and set proficiency levels on grade 11 writing for the State of Illinois. Cook has extensive facilitation experience, a variety of teaching experience, and experience with NAEP as the project director for the development of the 2011 NAEP Writing Framework. Allen had direct responsibility for psychometric aspects of NAEP while at Educational Testing Service. She also has a variety of experience teaching and conducting standard-setting meetings. All three have extensive experience in the delivery of professional presentations at professional and/or trade conferences.

Richard Duschl served as the content facilitator for grade 12, Senta Raizen was the content facilitator for grade 8, and Alice Fu was the content facilitator for grade 4. All three content facilitators were members of the 2009 NAEP Science Framework Planning Committee and Raizen was the co-chair of that committee. In addition, Raizen was the chair of the committee that developed the science ALDs, and Duschl was a member of that committee.

Because the meeting involves three grades, three sets of facilitators, and three groups of panelists, one large meeting room was used for joint grade level activities and separate smaller meeting rooms were used for individual grade level activities. Figure 1 illustrates a typical grade level room set

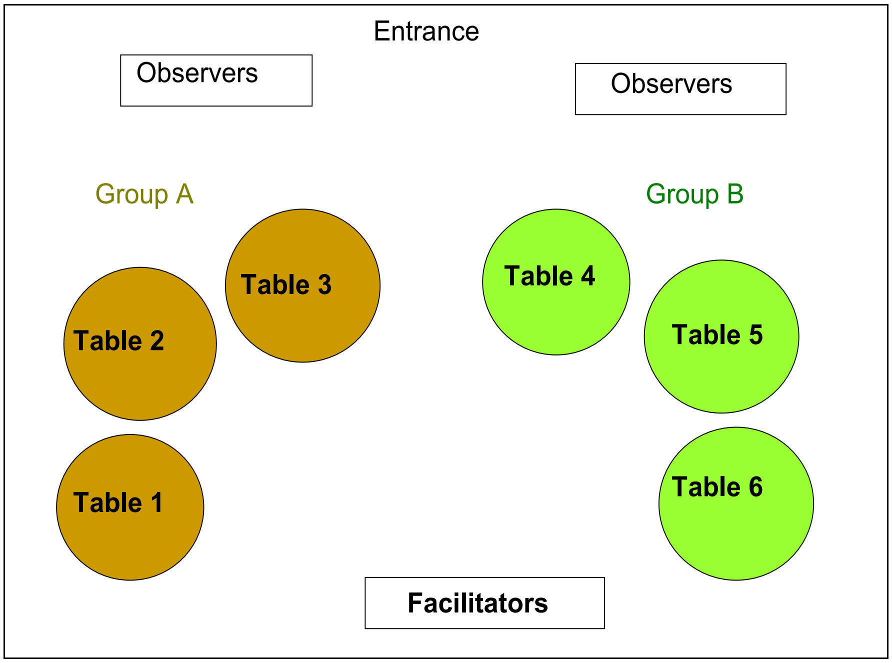


Figure 1: Grade level room and table setup

A total of six observers were present at various times: Susan Loomis, Assistant Director of Psychometrics at the Governing Board and contracting officer's representative (COR) for the contract; Andrew Kolstad, Senior Technical Advisor at National Center for Education Statistics; and, TACSS members Audrey Champagne, Professor Emerita from University of Albany, SUNY; Barbara Dodd, Professor from the University of Texas, Austin; Robert Forsyth, Professor Emeritus from University of Iowa; and Mary Pitoniak, Lead Program Administrator from Educational Testing Service. Observers were instructed not to participate in the process, not to signal approval or disapproval in any fashion that could be seen by the panelists, and not to distract the facilitators.

Orientation

The ALS meeting started with a welcome by Board member Senator van de Putte from Texas. Her welcome was followed by a series of whole-group orientation sessions attended by the panelists from all three grades.

Welcome and Introductions

In a brief welcome and introduction session, Nancy Petersen, the project director, introduced ACT staff, the process and content facilitators, and the observers to the panelists. The role of observers was explained and panelists were asked to limit their interactions with observers to matters not directly related to the process. In addition, the process for selecting panelists was described.

General Orientation to the NAEP

Following the welcome and introductions, Susan Loomis provided panelists with background information on NAEP and the Governing Board. This session covered the history, organizational structure, procedures, and key policies of the NAEP as well as the purpose of setting achievement levels. Achievement level percentages for the 2005 NAEP science assessments were also presented.

Taking and Scoring a NAEP Exam

After this general orientation, the panelists took a form of the 2009 science assessment for the grade level to which they were assigned. The test was administered under conditions similar to those followed for the actual student administration. Upon completion of the test, panelists were given scoring rubrics and scored their own performance. The test form administered was composed of the two blocks tentatively scheduled for release. These same two blocks also comprised the items common to both item-rating pools within a grade. Panelists were told that their test would not be scored or used in any other way during the meeting, but that they were to use the experience to gain some additional insight into what students experience when taking the test. This was also an opportunity for panelists to become familiar with the assessment items and scoring rubrics for the common items.

After completing the test, panelists were given training in how to use the scoring rubrics for CR items. Many of the CR items in the science assessments are not straight-forward (e.g., some CR items have multiple parts which were scored separately, summed, and collapsed to obtain the final score on the item). Following training on the rubrics, panelists were provided with scoring guides and given time to score their own responses.

The NAEP Science Framework

After panelists finished scoring their own tests, Duschl, the grade 12 content facilitator, provided panelists with an orientation to the Science Framework. Understanding the assessment framework is the first step panelists need to take toward reaching a useful understanding of what students in grades 4, 8, and 12 should know and be able to do in science at each achievement level. During the presentation, the purpose of the Science Framework was explained, the development process was described, and the science content and science practices measured in the assessment were reviewed. It was pointed out that the framework had the same basic structure for all three grades (see Table 16).

Table 16: 2009 NAEP science content topics and subtopics

Physical Science	Life Science	Earth and Space Sciences
Matter • Properties of matter • Changes in matter **Energy** • Forms of energy • Energy transfer and conservation **Motion** • Motion at the macroscopic level • Forces affecting motion	**Structures and Functions of Living Systems** • Organization and development • Matter and energy transformations • Interdependence **Changes in Living Systems** • Heredity and reproduction • Evolution and diversity	**Earth in Space and Time** • Objects in the universe • History of Earth **Earth Structures** • Properties of Earth materials • Tectonics **Earth Systems** • Energy in Earth systems • Climate and weather • Biogeochemical cycles

A main objective of this presentation was to show panelists how familiarity with the Science Framework would be helpful when they performed the item review task. Panelists were informed that the assessment items were written at the content topic level, not the subtopic level; and, when the panelists reviewed the items, they would be provided with the content area and the code for the content statement (see Table 17) associated with each item.

Table 17: Example of 2009 NAEP earth and space sciences content statements

Grade 4	Grade 8	Grade 12
Earth Systems		
Climate and Weather: Local weather (4), global weather patterns (8), systems that influence climate (12)		
E4.8: Weather changes from day to day and during the seasons. **E4.9:** Scientists use tools for observing, recording, and predicting weather changes from day to day and during the seasons.	**E8.13:** Global patterns of atmospheric movement influence local weather. Oceans have a major effect on climate because water in the oceans holds a large amount of heat.	**E12.10:** Climate is determined by energy transfer from the Sun at and near Earth's surface. This energy transfer is influenced by dynamic processes such as cloud cover, atmospheric gases, and Earth's rotation, as well as static conditions such as the positions of mountain ranges, oceans, seas, and lakes.

Panelists were told that items (or performance expectations) are derived from the crossing of content statements and science practices. They were informed that when they reviewed items they would need to create notes for each item describing "what students need to know and be able to do" to answer the item correctly and that their notes should use verbs to describe the skills students need to use to answer the item correctly. It was explained that they would find the general performance expectations for science practices to be a good source for appropriate verbs. They were told that during the item review task, they would be provided with the science practices (see Figure 2) associated with each item and a more detailed description of the general performance expectations for science practices from the Science Framework (see Exhibit 13 in NAGB, 2008c).

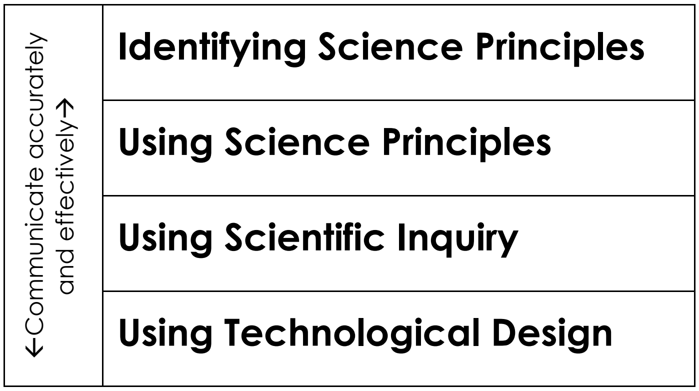

←Communicate accurately and effectively→
Identifying Science Principles
Using Science Principles
Using Scientific Inquiry
Using Technological Design

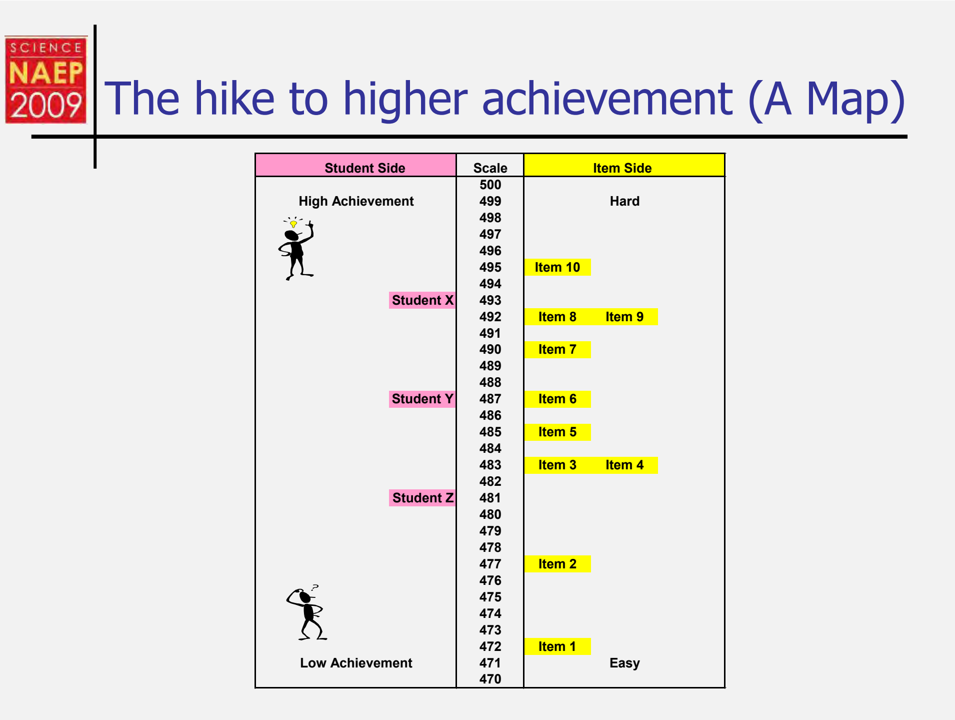
Figure 2: 2009 NAEP science practices

Orientation to the ALS Method

Once oriented to the NAEP, panelists learned about the Mapmark with Whole Booklet Feedback method in a whole-group orientation session given by Hamme Peterson, the grade 12 process facilitator. The purpose of this orientation was to explain how the ALS meeting fits into the overall NAEP assessment process and to describe some basic concepts and procedures that are central to the method.

In this session, Hamme Peterson first gave a general overview of the key steps in the NAEP assessment process, starting with framework development and ending with reporting of assessment results; and, she explained where the ALS meeting fit into this overall process. She then told panelists that their task for the meeting was to recommend cut scores for Basic, Proficient, and Advanced performance on the 2009 Science NAEP and to identify items that exemplify performance at those three achievement levels. She listed the steps they would follow to accomplish those tasks. It was pointed out that they had already completed the first step, which was to review the framework, and that the next step would be to review the items.

To prepare panelists for the item review task, information on the number and types of items on each assessment was provided. Training was then given on key materials and concepts used during item review (i.e., item map, RP criterion, and the OIB). Panelists were first introduced to the item map. Figure 3 shows a slide of a simplified item map. This slide was used to explain the general principle of an item map as spatially representative of a journey. In this case, the map represents the journey from low to high achievement. Points along the journey are represented by *landmarks*, i.e., test items.



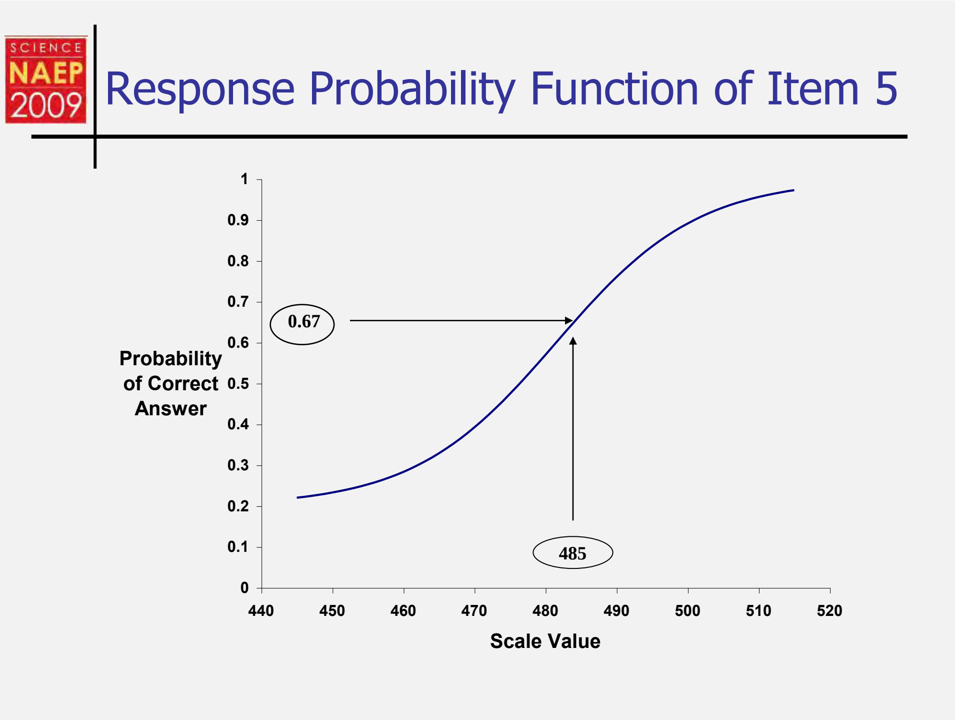
SCIENCE NAEP 2009

The hike to higher achievement (A Map)

Student Side	Scale	Item Side	
	500		
High Achievement	499	Hard	
	498		
	497		
	496		
	495	Item 10	
	494		
Student X	493		
	492	Item 8	Item 9
	491		
	490	Item 7	
	489		
	488		
Student Y	487	Item 6	
	486		
	485	Item 5	
	484		
	483	Item 3	Item 4
	482		
Student Z	481		
	480		
	479		
	478		
	477	Item 2	
	476		
	475		
	474		
	473		
	472	Item 1	
Low Achievement	471	Easy	
	470		

Figure 3: A simplified item map as spatially representative of a journey from low to high achievement

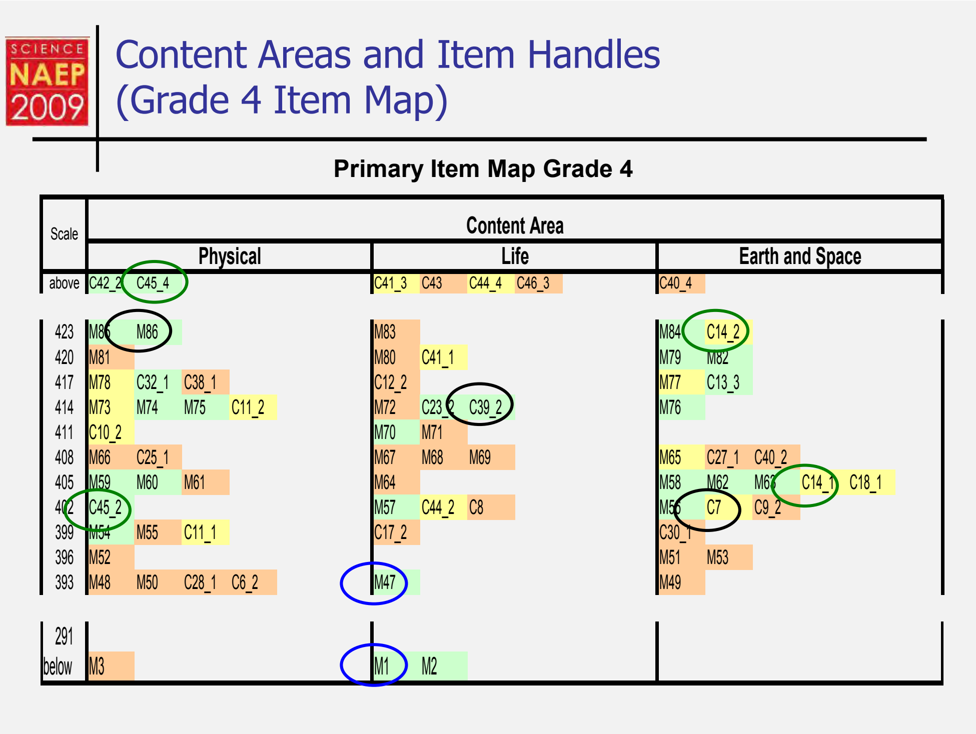
Figure 4 shows a slide used to explain the role of the RP criterion in determining the location of item landmarks on the map. This slide was used to explain the location of Item 5 in Figure 3 as a function of the probability of a correct answer on that item at a given score point on the assessment. Items were mapped to the assessment scale values based on an RP criterion of 0.67. In other words, an item was mapped to the scale value at which a student has a 0.67 (or 67%) chance of answering the item correctly. This RP criterion was used to define *mastery* and panelists were instructed to consider a 2-in-3 chance as meaning mastery of the relevant content reflected in the item. Introducing this concept early is important in helping panelists to understand this criterion and to take it into account in their bookmark placements.



Response Probability Function of Item 5

Figure 4: The relationship of the RP criterion to an item's scale value

Panelists were then shown the Primary Item Map (Appendix B), on which columns correspond to the content areas of the assessment. The item map illustrates the distribution of all of the assessment items on the achievement scale, mapped from easiest to hardest. Panelists were shown how this map would allow them to compare differences in difficulty between items by identifying the distance between those items on the map. The slide in Figure 5 showing part of an item map was used to explain the meaning of colors and other information in the Primary Item Map.



SCIENCE NAEP 2009

Content Areas and Item Handles (Grade 4 Item Map)

Primary Item Map Grade 4

Scale	Content Area		
	Physical	Life	Earth and Space
above	C42_2 C45_4	C41_3 C43 C44_4 C46_3	C40_4
423	M85 M86	M83	M84 C14_2
420	M81	M80 C41_1	M79 M82
417	M78 C32_1 C38_1	C12_2	M77 C13_3
414	M73 M74 M75 C11_2	M72 C23_2 C39_2	M76
411	C10_2	M70 M71	
408	M66 C25_1	M67 M68 M69	M65 C27_1 C40_2
405	M59 M60 M61	M64	M58 M62 M63 C14_1 C18_1
402	C45_2	M57 C44_2 C8	M56 C7 C9_2
399	M54 M55 C11_1	C17_2	C30_1
396	M52		M51 M53
393	M48 M50 C28_1 C6_2	M47	M49
291			
below	M3	M1 M2	

Figure 5: Illustration of how items are displayed on an item map

Part of an item map is shown in Figure 5. The item map shown is from Grade 4, but Grade 8 and 12 maps are very similar, only the score values and item handles differ. Separating the items into content related columns (Physical, Life, and Earth and Space) provides the panelists with a layer of organization when they look at the map. This allows them to see which items measure a related set of skills (skills within a content area) and to think about what makes one item more difficult than another within a content area. To make the item maps a manageable size, score intervals on the item maps were three scale points wide. The item map for each grade was printed as an 8½ x 11 inch document.

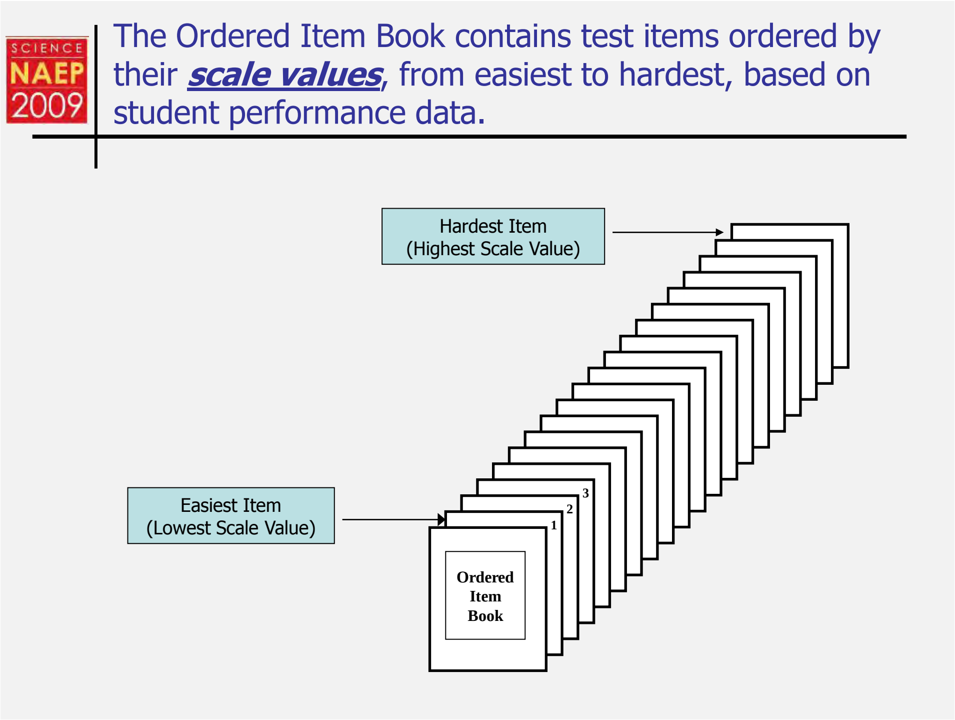
Each item is represented on the map by a handle—a unique identifier—consisting of a character followed by a number (e.g., M1, C1, C39_2). The first digit of the handle represents item type (C = constructed response and M = multiple choice). The number following the character represents where that item falls in order of difficulty within type. For example, M1 is the easiest MC item and C7 is the 7th most difficult CR item on the grade 4 science assessment. The difficulty rank of each item is based on the difficulty of receiving full credit (the last or highest score level) on the item.

The scoring of extended CR items allows for partial credit. For example on a two-point extended CR item, a student whose response is partially correct will get one point and a student whose response is fully correct will get two points, or full credit. Extended CR items occur in multiple places on the item map, one place for each possible score level. Handles for extended CR items include an underscore "_" followed by the score level. Short CR (or dichotomous) items only have one score level so their handle does not include a dash and number. C7 is an example of a dichotomously-scored CR item.

The score locations of C14, a two-point CR item, are circled on the map in Figure 5. The scale value for the first score point, C14_1, is in the map score interval with midpoint 405; and, the scale value of the second score point, C14_2, is in the score interval with midpoint 423.

The color of an item handle on the map indicates whether the item is in the Group A pool only (tan), the Group B pool only (green), or in both item pools (yellow). Item C14 was in both item pools. Items in both pools are *common* items.

Panelists were then oriented to the OIB, which accompanied the Primary Item Map. The OIB contained all of the items with which the panelists would be working in order of their difficulty, beginning with the easiest. Figure 6 shows a slide used to illustrate this concept.



SCIENCE NAEP 2009

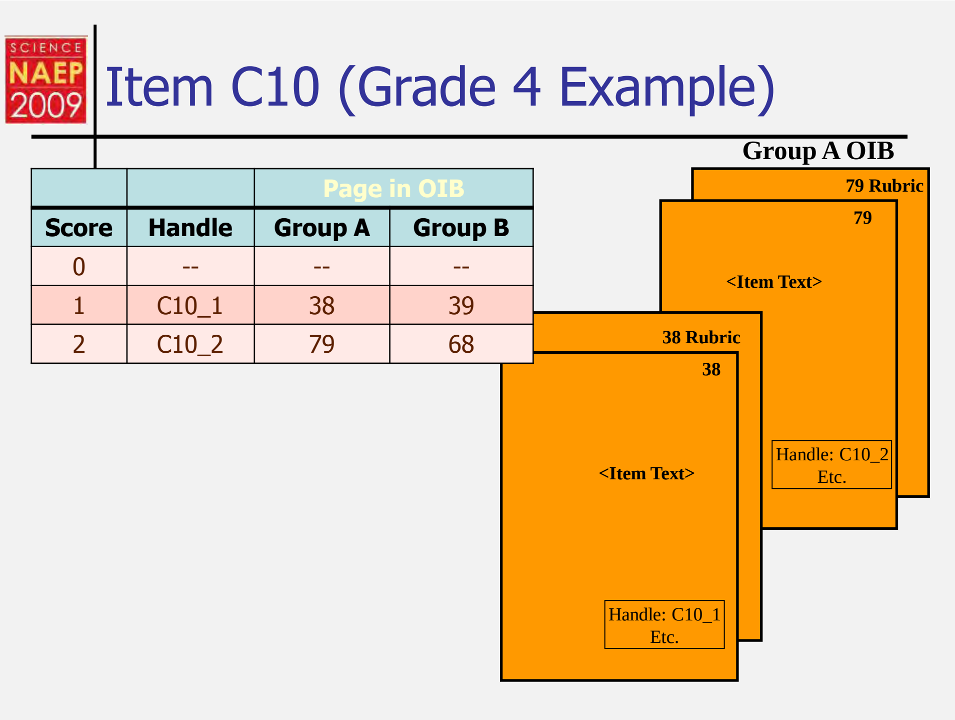
The Ordered Item Book contains test items ordered by their ***scale values***, from easiest to hardest, based on student performance data.

Hardest Item
(Highest Scale Value)

Easiest Item
(Lowest Scale Value)

Figure 6: Illustration of how items are ordered by difficulty in the Ordered Item Book (OIB)

The slide in Figure 7 shows the location of the two score points of item C10 in the Group A and Group B OIBs and indicates the information contained in the OIB for each score point. Score points of extended CR items were treated as separate items in the OIB, just as they were on the item map. In the Group A OIB, the first score point of item C10 was located on page 38 and the second score point was located on page 79. There were at least two pages for each score point of a CR item in the OIB—one showing the item and one showing the scoring rubric—but the page numbers in the OIB increased only when the item or score level changed.



SCIENCE NAEP 2009

Item C10 (Grade 4 Example)

		Page in OIB	
Score	Handle	Group A	Group B
0	--	--	--
1	C10_1	38	39
2	C10_2	79	68

Group A OIB

79 Rubric

79

<Item Text>

Handle: C10_2
Etc.

38 Rubric

38

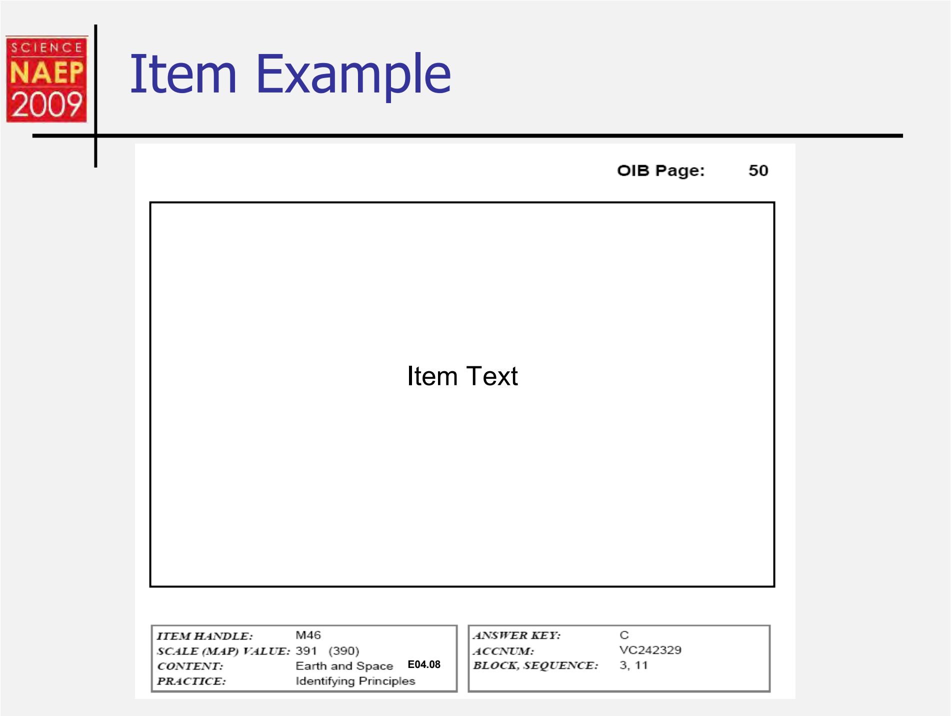
<Item Text>

Handle: C10_1
Etc.

Figure 7: Illustration showing item location and information location for Item C10 in the OIB

On the OIB page that contained the item's text, there was a framed box, as shown in Figure 8. The information box was brought to panelists' attention and the information was explained. The box contained the item's or score-point's:

- handle,
- scale value (the scale value at which a student has a 0.67 probability of earning the score point or correctly answering the item),
- map value (the midpoint of the interval containing the item on the item map),
- content area and specific content statement classification in the 2009 Science Framework associated with the item,
- science practice classification in the 2009 Science Framework associated with the item,
- answer key,
- identification code, and
- block and sequence number.



SCIENCE NAEP 2009

Item Example

OIB Page: 50

Item Text

ITEM HANDLE: M46
SCALE (MAP) VALUE: 391 (390)
CONTENT: Earth and Space E04.08
PRACTICE: Identifying Principles

ANSWER KEY: C
ACCNUM: VC242329
BLOCK, SEQUENCE: 3, 11

Figure 8: Illustration of the information on an OIB page

Following the orientation to key materials and concepts, panelists were told that after the completion of item review on day 2, they would review the ALDs. The ALDs describe student performance at three achievement levels—Basic, Proficient, and Advanced. It was explained that the role of panelists as standard setters is to establish cut scores on the assessment that reflect the content of the ALDs. In other words, a student who scores at a proficient level *should* be able to complete items on the assessment that reflect the proficient level description. Panelists were told that as they worked through the exercises over the next few days, they would be attempting to ascertain what material from the assessment a student needs to have mastered to *just* meet the description of each achievement level so as to determine what scale score best reflects *borderline* Basic, Proficient, and Advanced performance.

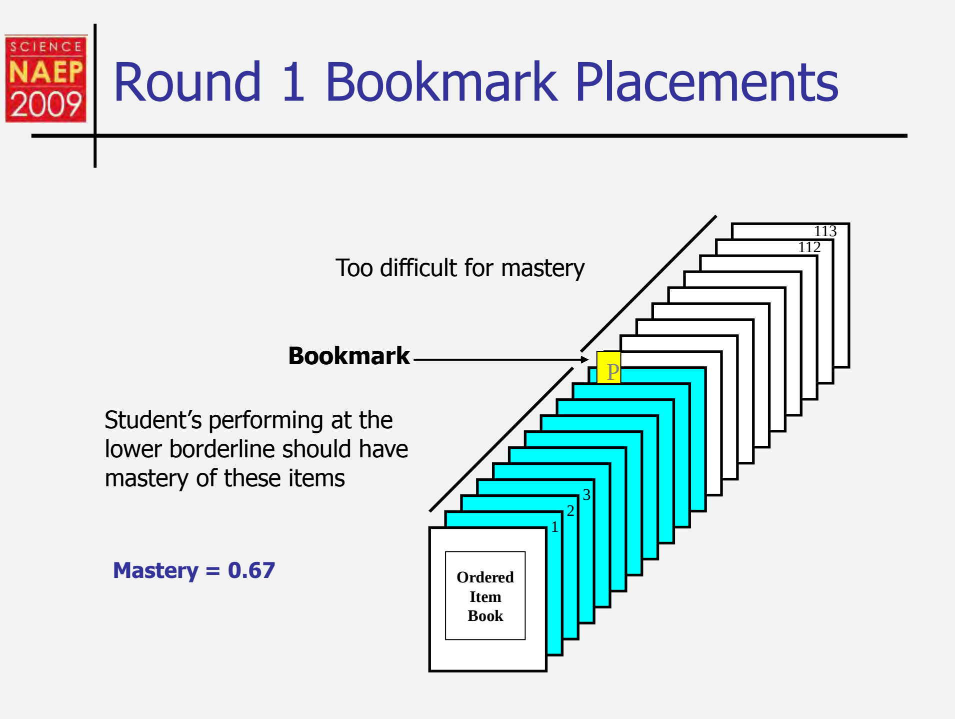
The Governing Board's general policy definitions for Basic, Proficient, and Advanced performance were presented (NAGB, 2008b). The Governing Board also requires that ALDs specific to each subject area be written for each achievement level for each grade. The subject-specific ALDs:

- state what students should know and be able to do in science at Basic, Proficient, and Advanced levels in each of the three grades,
- define outcomes of 4th, 8th, and 12th grade science education, and
- provide clarity and meaning to assessment results.

The ALDs were created by the Governing Board and an ALD committee comprised of science experts outside of the scope of this standard-setting project. Panelists used the subject-specific ALDs (Appendix C) to set their cut scores.

After the introduction to the ALDs, panelists were given a high-level overview of the Mapmark with Whole Booklet Feedback method. They were told it is a three round process. They were given a brief description of the key activities in each round. It was explained that each round was designed to provide them with information and feedback to set their cut score in round 1 and then to evaluate and possibly change that cut score in successive rounds. They were told that the cut scores emerging from round 3 would be the final cut scores. They would not set cut scores after round 3, but they would complete a questionnaire indicating the degree to which they felt comfortable with the final cut scores. It was explained that their final task would be to help ACT make recommendations to the Governing Board as to which items on the assessment best exemplify performance at each of the three achievement levels. They would be provided with *potential* exemplar items drawn from test blocks that the Governing Board may release to the public. In this task, they would rate the potential exemplar items on whether they felt they should be used to illustrate what students in a particular achievement level should know and be able to do. They were informed that before and during each round and during the exemplar item rating, their grade level process facilitator would provide detailed instructions for each task.

Next, the slide in Figure 9 was used to briefly describe the bookmark placement process for setting cut scores in round 1. It was pointed out that panelists would place three bookmarks by the conclusion of round 1—one for each of the achievement levels—Basic, Proficient, and Advanced. Time was spent explaining that the setting of cut scores would be criterion-referenced, based on the ALDs, rather than norm-referenced.



SCIENCE NAEP 2009

Round 1 Bookmark Placements

Too difficult for mastery

Bookmark

Student's performing at the lower borderline should have mastery of these items

Mastery = 0.67

Figure 9: Illustration of round 1 bookmark placement

The orientation to the ALS method session concluded with a quick overview of the meeting agenda. Panelists were then dispersed to the rooms for their grade groups to start the task of setting cut scores for each achievement level.

Round 1: Understanding the Assessment and Student Achievement

Overview of Round 1

The first round of the ALS method included a review of the items with identification of what a student needs to know and be able to do to get a MC item correct or to score points on a CR item; a presentation of the ALDs; and the placement of the bookmark for each achievement level. By the end of round 1, panelists had set a bookmark for each achievement level—Basic, Proficient, and Advanced.

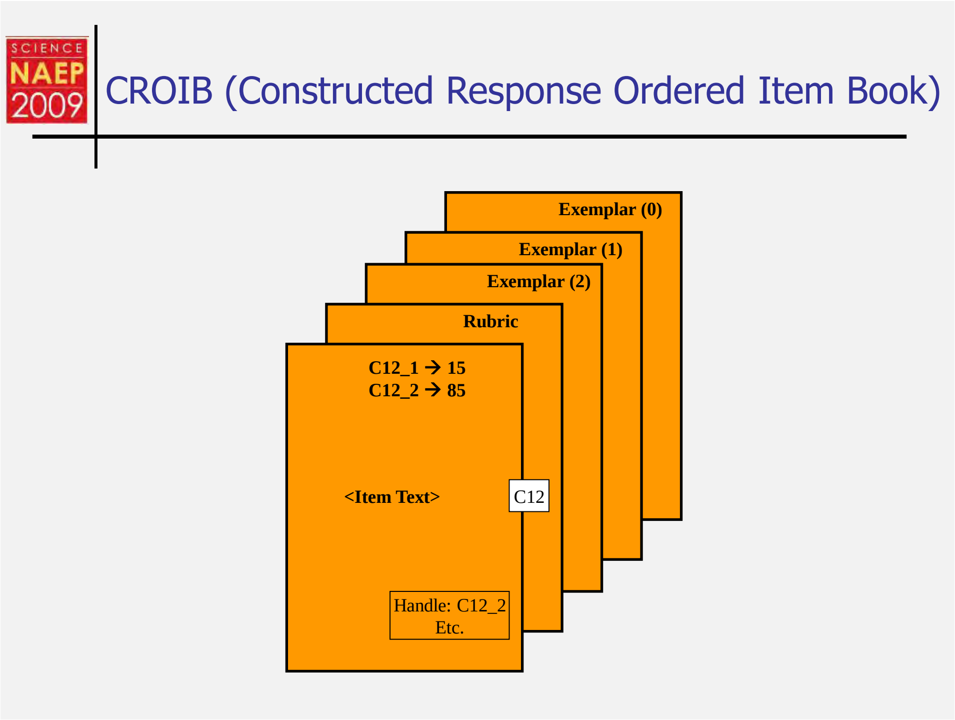
Item Review

To set cut scores on an assessment, panelists must have a good understanding of the assessment and the knowledge and skills the assessment requires students to demonstrate in order to earn successively higher scores on the test. Panelists spent eight and a half hours of meeting time identifying the science content and practices that students need to know and be able to use in order to earn full credit on successively more difficult items on the test. There were four stages to this activity.

> Stage 1—Grade-group review of selected common CR items. This was a grade-group discussion, led by the content and process facilitators for each grade, in which panelists were trained in the process of identifying the science content and practices required by CR items. The content and process facilitators modeled the item review task for four items common to both item pools that illustrated the various types of scoring rubrics associated with the CR items. They began with an easy item and proceeded to look at increasingly more difficult items. For each CR item, they first identified and made notes on what students needed to know and be able to do to get full credit on the item; then they identified and made notes on the knowledge and skills needed to earn successively lower scores on the item. Maximum scores on the CR items ranged from 1 to 4 points per question.
>
> Stage 2—Table-group review of remaining CR items. This item review task was conducted in table groups. Each table applied the process modeled in stage 1 to the remaining CR items in their item-rating pool. Panelists took turns leading this activity at their table. Panelists discussed what science content and practices students need to know and be able to do to get full credit on the item. Following group discussion of each item, each panelist made notes as to the knowledge and skills they thought were needed to earn successively lower scores on the item. Content and process facilitators circulated among the tables, providing guidance on the process, and answering questions as needed.
>
> Stage 3—Independent review of OIB. This was an independent item review task in which panelists identified the knowledge and skills required by all of the items in their pool in the context of their OIB. They considered items sequentially, beginning with the first, or easiest, item. During the independent review of the OIB, panelists made notes on what students need to know and be able to do to answer each MC item correctly and they transferred their notes on CR score points to the OIB as they encountered each score point. This review enabled panelists to become familiar with the progression of difficulty from one item to the next within their item-rating pool.
>
> Stage 4—Table group discussion of OIB. This was a table-group discussion of the science content and practices associated with each item/score point in the context of the OIB. Again, items were considered sequentially, beginning with the easiest. Panelists shared their thoughts as to the knowledge and skills needed to get each item correct and added to their notes ideas they agreed with but had not previously noted.

Materials for stages 1 and 2 of the item review included the CROIB and a Notes template. The CROIB contained all the CR items in a group's item pool. Items were listed in order of difficulty by the highest score point.

The slide in Figure 10 illustrates the contents of the CROIB. Unlike the OIB, all the information about a CR item was contained together, on consecutive pages within the CROIB. Items were separated by tabbed pages, with the tab showing the item handle (minus the score points). Item information included the scoring rubric and examples of student responses at each score level, including zero. The first page showed the item, the information box, and the page number(s) where the item's score point(s) could be found in the OIB.



SCIENCE NAEP 2009

CROIB (Constructed Response Ordered Item Book)

Exemplar (0)

Exemplar (1)

Exemplar (2)

Rubric

C12_1 → 15
C12_2 → 85

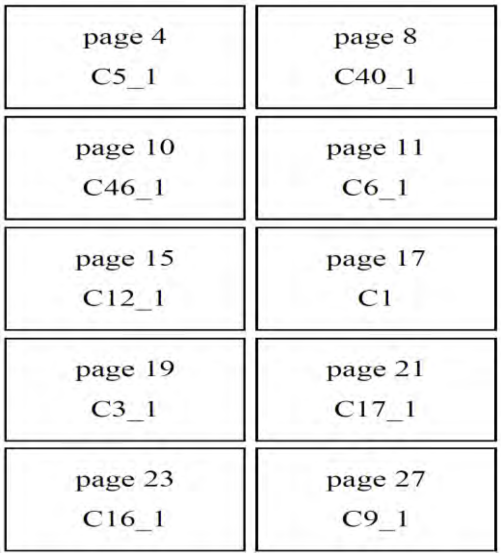
<Item Text>

C12

Handle: C12_2
Etc.

Figure 10: Illustration showing information location for Item C12 in the CROIB

Panelists used large yellow Post-It Notes to record their notes on the science content and practices associated with each CR score point. They were told that their notes were for their own use. They used one Post-It Note for each score point. When panelists were finished with an item, they placed their notes in the Notes template (see Figure 11). This was a stapled set of 11 x 17 pages with outlines for accommodating ten Post-It Notes per page. Within each Post-It outline was an item handle and OIB page number identifying the Post-It Note that was to be placed there.



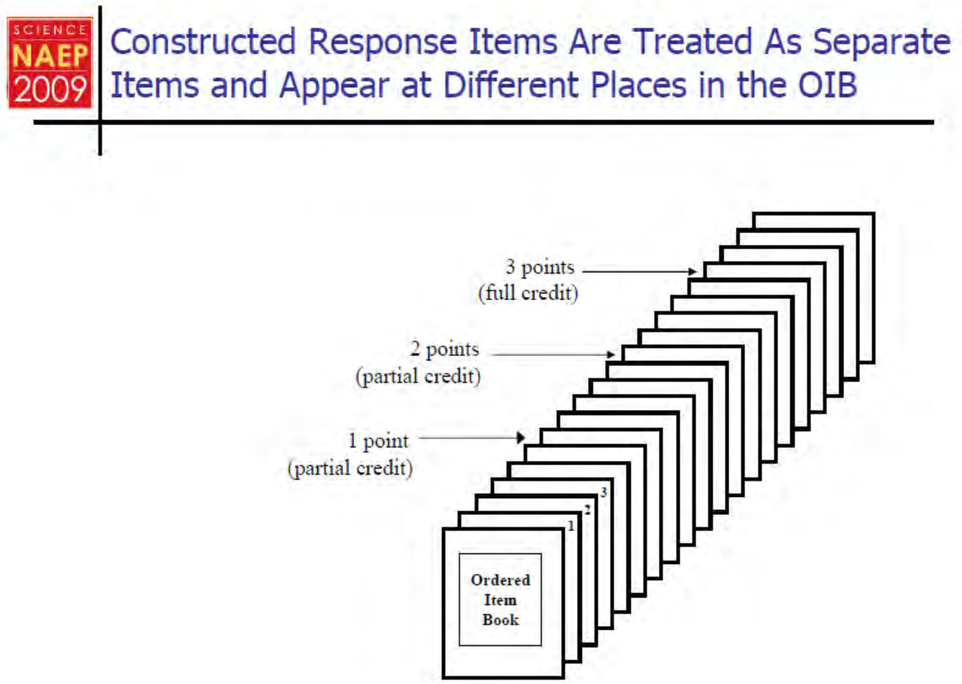
page 4 C5_1	page 8 C40_1
page 10 C46_1	page 11 C6_1
page 15 C12_1	page 17 C1
page 19 C3_1	page 21 C17_1
page 23 C16_1	page 27 C9_1

Figure 11: Illustration of a Notes template page

During stage 3, panelists made notes on what students need to know and be able to do to get the MC items correct. Because of time constraints, each panelist reviewed and made notes on only 80% of the MC items in the OIB. To ensure that every MC item was reviewed by at least one panelist at each table, each panelist was given a list of specific MC items to review.

As panelists progressed through their OIB, they transferred their notes on CR score points from the Notes template to the corresponding OIB page as they encountered each score point. As noted earlier, the OIB contained all items, including the CR response items. Figure 12 shows how score levels of extended CR items were treated as separate items in the OIB. The use of the Notes template allowed panelists to place their notes on the scored item steps on the correct OIB page with just one pass through the OIB. This allowed panelists to see their CR item notes in the context of all of the items in the OIB. When panelists saw score points of extended CR items relative to the difficulty of all other items in their pool, they could add to their notes observations about what science content and practices the score point may require that previous, easier items and score points did not require. Panelists recorded further notes directly on the pages of the OIB.

Panelists also checked MC and CR items off on their Primary Item Map as they progressed through the OIB. The item check-off process helped panelists see “how much” more difficult one item was than another and which items were related in terms of the general knowledge and skills that distinguished different content areas.



SCIENCE NAEP 2009

Constructed Response Items Are Treated As Separate Items and Appear at Different Places in the OIB

Figure 12: Illustration showing the location of extended CR item score levels in the OIB

During the table-group discussion in stage 4, panelists shared their ideas about what students needed to know and be able to do and added the ideas of other panelists to their notes. Panelists took turns leading the table discussion. The process was monitored by facilitators to reinforce the idea that all panelists have something valuable to contribute to the process.

When the item review was complete, panelists had a detailed, *structured* understanding of the assessment and student achievement. Structure was provided by the difficulty-order of the knowledge and skills required by test items as shown in the OIB and on the Primary Item Map. This structure prepared panelists to understand the continuum of increasing knowledge and skills represented by the Basic, Proficient, and Advanced ALDs.

Understanding the ALDs

Following completion of the item review task, the content facilitators led grade-group discussions of the ALDs. First, they reviewed the purpose, meaning, and structure of the ALDs. Then, they led the panelists in two activities. The first activity was designed to help panelists become familiar with the ALDs for their grade level and the progression of expectations across achievement levels. Panelists were asked "What differences do you see as the levels move from Basic to Proficient to Advanced?" They were told to read the descriptions of the science practices under each of the three achievement levels and underline key terms, think about the progression in science practices represented by the three levels, and share their observations about progression of expectations across the three levels within their table group. The content facilitator then called for some volunteers to share their observations and the process facilitator wrote the observations reported on a flip chart. Next, panelists were asked to consider the differences across levels in relation to the science principles, and then in relation to the achievement level summaries.

The second activity was designed to help panelists focus more directly on achievement at the borderline of each achievement level. This activity was done in table groups with table representatives reporting on the discussion. First, panelists were asked to describe the minimal achievement a student must have to be considered Proficient. Next, panelists were asked to describe the minimal achievement a student must have to be considered Basic, then Advanced. When panelists shared their observations about borderline achievement for each level, the process facilitator wrote the key terms from the reports on a flip chart for later reference.

Finally, the content facilitator briefly related performance on items to the ALDs to show panelists that while there is a relationship between achievement levels and item difficulty, there are wide ranges and overlaps. Plots were shared with the panelists showing that item difficulty is not closely related to content area, science practices, item type, or number of score points. Panelists were also shown items to illustrate that items testing the same content may be difficult or easy depending on the science practice required.

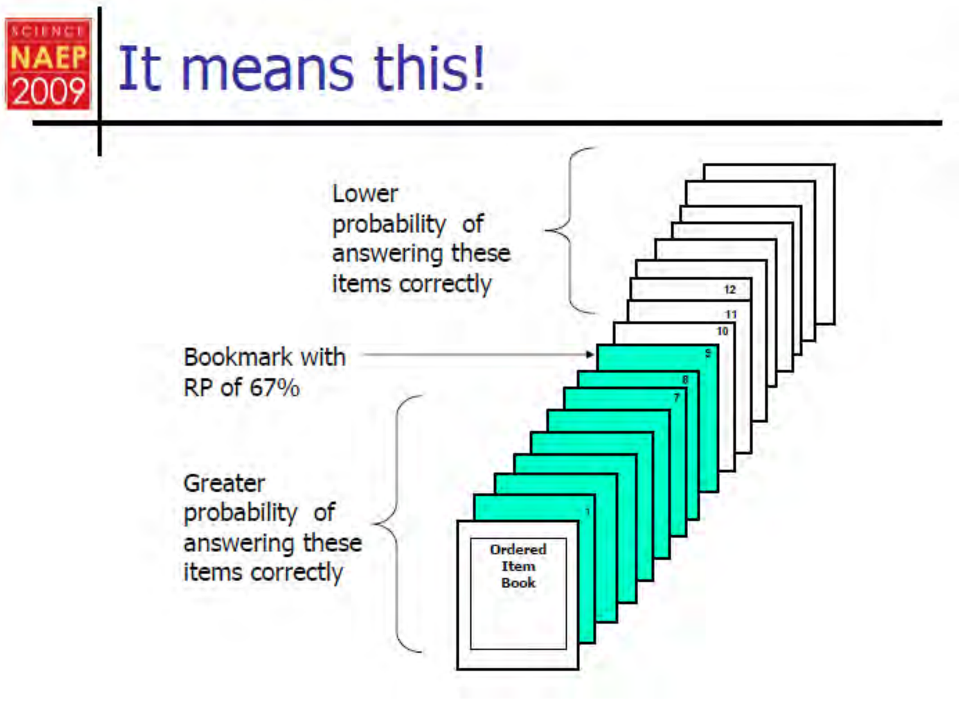
Placing the Bookmarks

The bookmark placement task began with a carefully scripted presentation that reemphasized the following points from the ALD discussion:

- Each ALD should be thought of as representing a *range* of performance on the achievement scale.
- The panelist's job is to decide what the lower *borderline* of that range should be.

Panelists were told to think of the lower borderline in terms of a student who was *just qualified* to be in the achievement level and to decide for themselves what *just qualified* meant in the process of placing their bookmarks. The structure provided by the OIB and Primary Item Map made it possible for panelists to develop and apply a concept of borderline in the process of placing their bookmarks.

The bookmark placement task was initially described to panelists as a process of going through the OIB, beginning with the easiest item, until they came to an item that they judged to be too difficult for mastery by the borderline student. Mastery was defined as having at least a 0.67 probability of answering the item correctly. The bookmark was placed on the item immediately preceding the too difficult item. The slide in Figure 13 was used to illustrate what is meant by mastery of the items at and below the cut score—mastery means that a student at the potential cut score has a 0.67 probability of answering the item at that cut score correctly, a higher probability of answering items below the cut score correctly, and a lower probability of answering the items above the cut score correctly.



SCIENCE NAEP 2009

It means this!

Lower probability of answering these items correctly

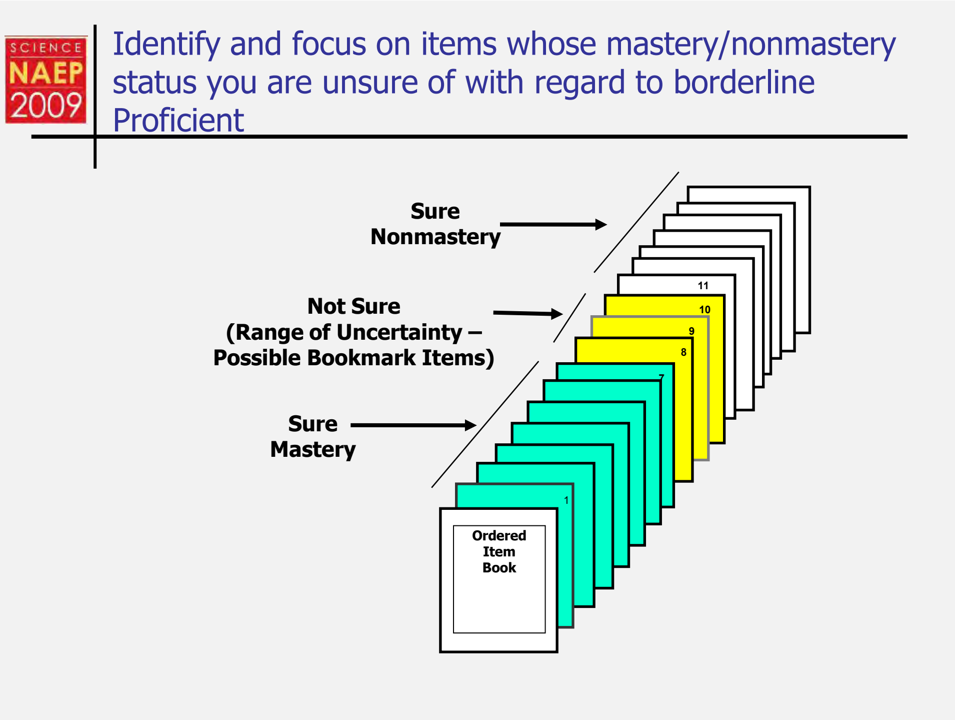
Bookmark with RP of 67%

Greater probability of answering these items correctly

Figure 13: Illustration of the relationship between bookmark placement and the mastery of items at and below the cut score using 0.67 as the RP criterion

Once panelists understood this idea, the facilitator explained to panelists that it was possible for them to be unsure of where to place their bookmarks because: (a) they may feel there was a noticeable or meaningful difference between adjacent items in terms of difficulty, and (b) they may feel that a few items in the OIB were out of order with their own expectations of relative difficulty.

The initial description of the process was then supplemented with the instruction to go beyond the first item they judge to be too difficult, to see if there were any later items that they felt the borderline student should have mastered. This instruction was represented to panelists visually by showing a *range of uncertainty* in a slide depiction of the OIB. All items below this range were *sure mastery* items. All items above this range were *sure nonmastery* items. Figure 14 shows a slide that was used to illustrate this concept for panelists.


SCIENCE NAEP 2009

Identify and focus on items whose mastery/nonmastery status you are unsure of with regard to borderline Proficient

Figure 14: Illustration of the range of uncertainty in bookmark placements

Bookmark placements were done one achievement level at a time starting with Proficient, then Basic, then Advanced. Panelists read the ALDs and used their understanding of the borderline of the ALD for the given level to place their bookmark for that level. Panelists were instructed to place their bookmarks independently, without discussion within their table group. After all panelists had placed their Proficient bookmarks, they were given instructions for placing their Basic and Advanced bookmarks individually, but at their own pace.

After placing all bookmarks, panelists were given an opportunity to adjust their bookmark placements. Panelists were encouraged to look at all of the ALDs together and to consider whether the differences between their bookmark placements were consistent with the increments of achievement implied by the ALDs. Finally, they were instructed to note the location of their bookmarked items on their item map.

Panelists recorded the page number of their bookmark placements on a Cut Score Recommendation form along with the page numbers corresponding to their range of uncertainty for each bookmark. They also circled the handles of their bookmarked items on their Primary Item Map. Staff then wrote the scale value corresponding to the bookmarked page beneath the bookmarked page number on each panelist’s Cut Score Recommendation form. The group cut score was computed by selecting the median cut score for each achievement level.

Round 2: Whole Booklet Feedback

Overview of Round 2

The second round of the ALS method started with presentation of the cut score results from round 1. Panelists then received holistic feedback in the form of actual student test booklets to help with their understanding of what students in the levels that they have now defined can do.

Following review and discussion of the student booklets, panelists selected a scale value representing their cut score for each achievement level.

Feedback from Round 1

Feedback from round 1 consisted of the median cut scores and the cut score distribution for the grade group from round 1. Figure 15 shows a cut score distribution chart provided as feedback from round 1. This chart was used to illustrate the location of all panelists' round 1 cut scores for each achievement level, the overlap (if any) between cut scores for achievement levels, and the highest and lowest cut scores by level.

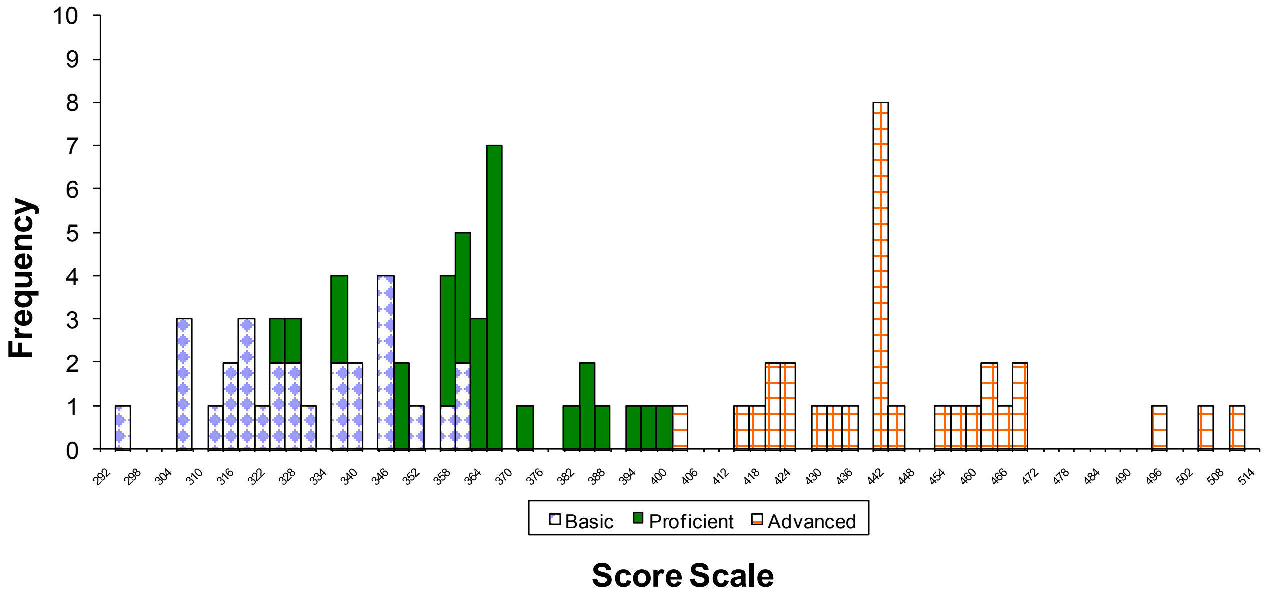
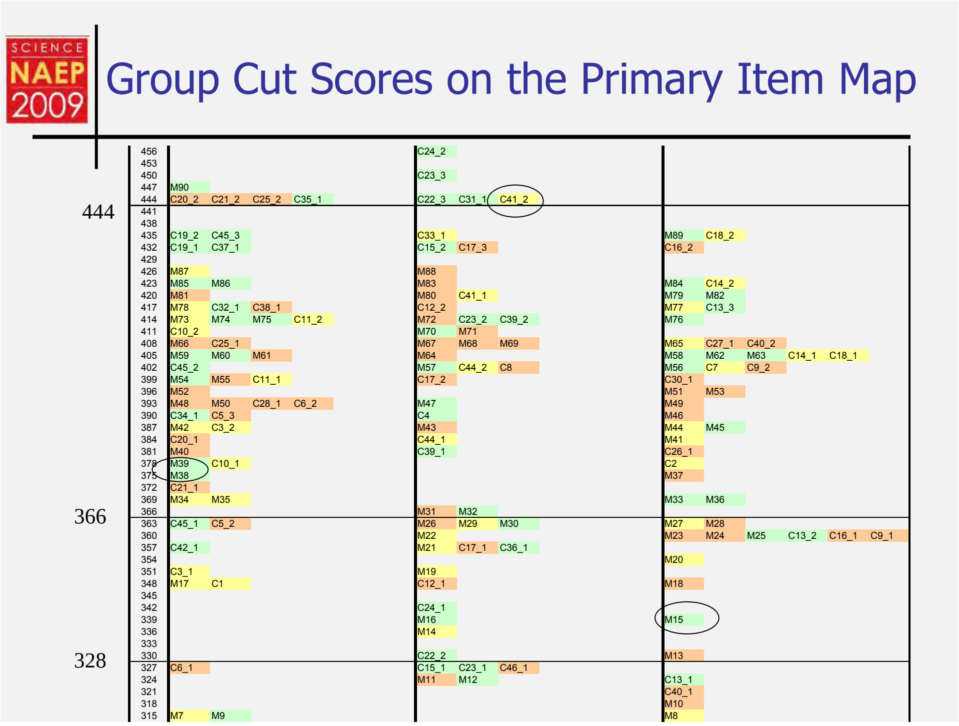


Figure 15: Cut score distribution chart showing the distribution of cut scores by achievement level

In addition to providing the numerical values of the cut scores, feedback was shown on item maps. Panelists were given a new version of their Primary Item Map with the group cut scores marked on the map as shown in Figure 16. Panelists then circled their round 1 bookmarked items on the item map so they could compare the group cut scores to their own cut scores.



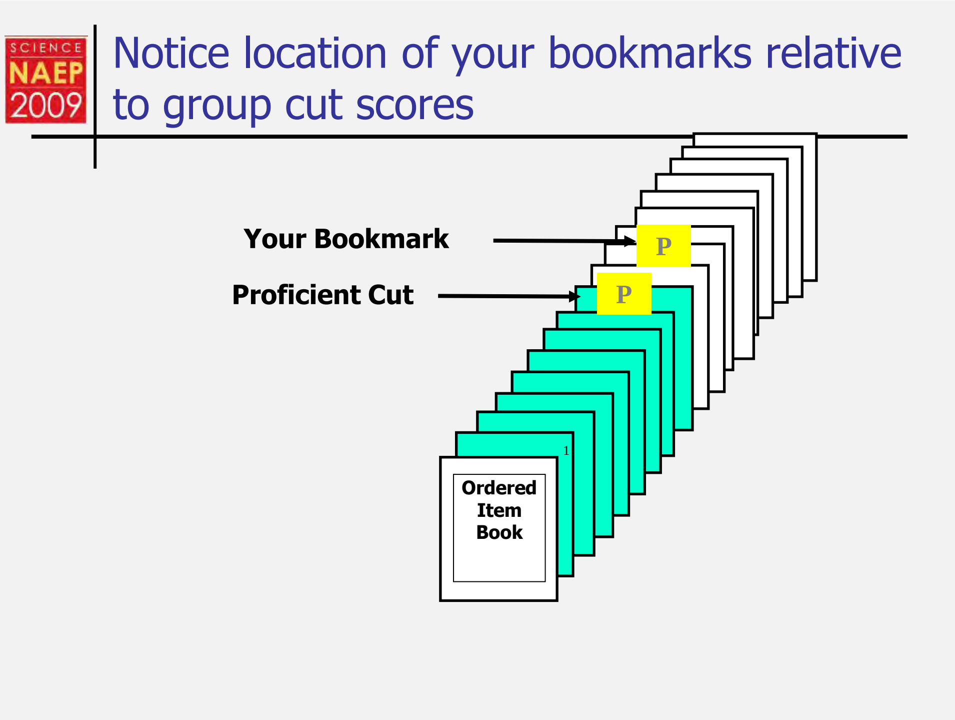
SCIENCE NAEP 2009

Group Cut Scores on the Primary Item Map

	456		C24_2	
	453			
	450		C23_3	
	447	M90		
	444	C20_2 C21_2 C25_2 C35_1	C22_3 C31_1 C41_2	
444	441			
	438			
	435	C19_2 C45_3	C33_1	M89 C18_2
	432	C19_1 C37_1	C15_2 C17_3	C16_2
	429			
	426	M87	M88	
	423	M85 M86	M83	M84 C14_2
	420	M81	M80 C41_1	M79 M82
	417	M78 C32_1 C38_1	C12_2	M77 C13_3
	414	M73 M74 M75 C11_2	M72 C23_2 C39_2	M76
	411	C10_2	M70 M71	
	408	M66 C25_1	M67 M68 M69	M65 C27_1 C40_2
	405	M59 M60 M61	M64	M58 M62 M63 C14_1 C18_1
	402	C45_2	M57 C44_2 C8	M56 C7 C9_2
	399	M54 M55 C11_1	C17_2	C30_1
	396	M52		M51 M53
	393	M48 M50 C28_1 C6_2	M47	M49
	390	C34_1 C5_3	C4	M46
	387	M42 C3_2	M43	M44 M45
	384	C20_1	C44_1	M41
	381	M40	C39_1	C26_1
	378	M39 C10_1		C2
	375	M38		M37
	372	C21_1		
	369	M34 M35		M33 M36
	366		M31 M32	
366	363	C45_1 C5_2	M26 M29 M30	M27 M28
	360		M22	M23 M24 M25 C13_2 C16_1 C9_1
	357	C42_1	M21 C17_1 C36_1	
	354			M20
	351	C3_1	M19	
	348	M17 C1	C12_1	M18
	345			
	342		C24_1	
	339		M16	M15
	336		M14	
	333			
	330		C22_2	M13
328	327	C6_1	C15_1 C23_1 C46_1	
	324		M11 M12	C13_1
	321			C40_1
	318			M10
	315	M7 M9		M8

Figure 16: Primary Item Map showing round 1 group cut scores (horizontal lines) and the location of a panelist's bookmarked items (circled)

Panelists were also instructed to place Post-it Notes on the group cut scores in their OIBs. To focus their attention on the intended, criterion-referenced meaning of the round 1 cut scores, panelists were instructed to identify, for each achievement level, the items that fell between their cut scores and the group's and to determine what these items represented in terms of differences in performance between the two definitions of borderline, as shown in Figure 17. They were instructed to keep in mind where their cut scores fell in relation to the group's, because examples of student performance would be provided at the group cut score and not the individual panelist's cut score.



Notice location of your bookmarks relative to group cut scores

Figure 17: Illustration of the comparison of the group cut score and a panelist's bookmarked item in the OIB.

Whole Booklet Feedback

Panelists were told that their round 2 cut score recommendations would incorporate judgments of whether performance exhibited in student booklets with scores at the borderline of each of the achievement levels was too low, OK, or too high for the borderline of that level. Booklets for ten students on each of three forms were provided, with each group (A and B) reviewing two forms for a total of 20 booklets per group. The booklets for each form were selected so that the student booklet scores were distributed across the achievement scale with two booklets at each grade-group cut score and one at the middle of each achievement level range, including Below Basic. For booklets within an achievement level, the booklets selected were the ones closest to the scale score that was at the midpoint of the achievement level. For the Advanced level, the scale score was the midpoint between the cut score for that level, and the scale score associated with the most difficult item. For the Below Basic level, the scale score was the midpoint between the cut score for the Basic level, and the scale score associated with the easiest item.

For each form, the expected number of points for each achievement scale value was plotted on the Booklet Score Plot, and the booklets were indicated on the plot at their scale value (see Figure 18). These plots were used to provide a visual illustration of the location of each booklet relative to the cut scores and the achievement scale.

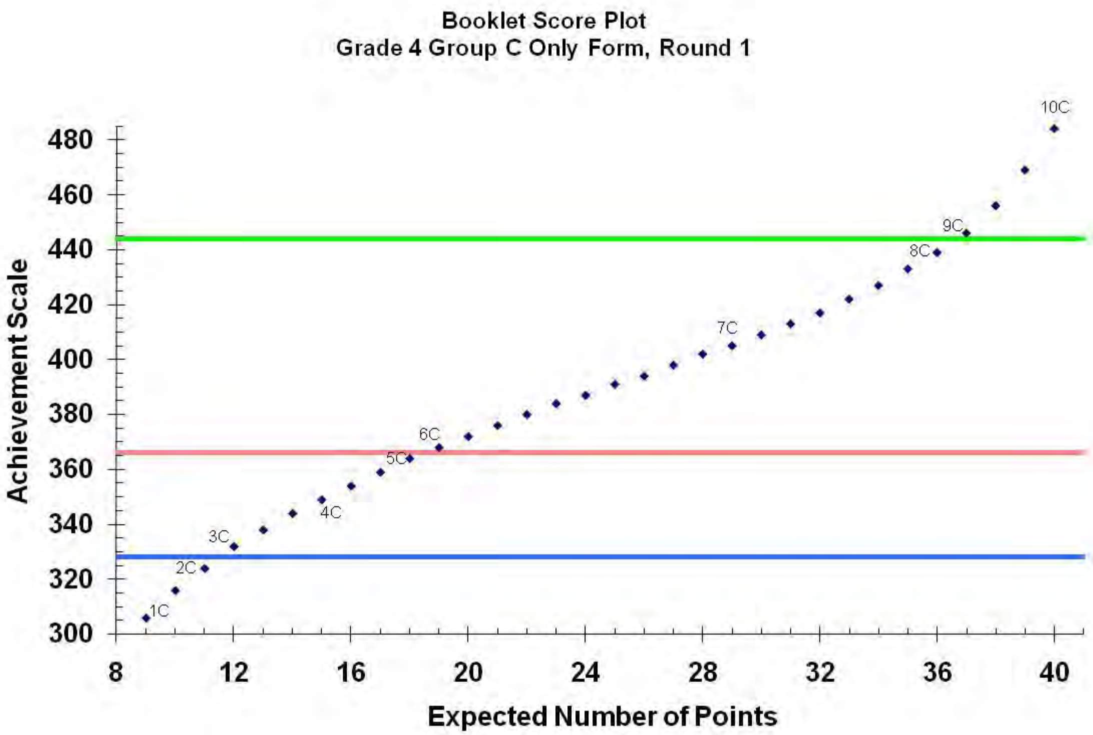


Figure 18: Booklet Score Plot for the common form showing the round 1 cut scores (horizontal lines) and the score of each student booklet (1C through 10C) on the achievement scale

Before panelists began their independent review of the student booklets, they were led through a grade-group exercise to familiarize them with the Booklet Score Chart (BSC), Item Score Table (IST), and booklet item maps and to help them begin to understand the relationship between general performance on a form of the test and expected performance on individual test items.

The BSCs were specific to each group and were provided for each achievement level. These charts mapped the expected number of points correct on the common and group-specific forms to the achievement scale within a range from 10 points below the *low* cut score for the achievement level to 10 points above the *high* cut score for the achievement level from round 1. The placement of the booklets on the chart was determined by their expected number of points correct. Panelists were asked to circle their cut scores on the BSC and to take note of where their cut scores fell in relation to the booklets they would be reviewing (see example in Figure 19).

Booklet Score Chart - Grade 4 Group A

Proficient

		Common Form		Group A Only Form	
	Scale	Booklet	Expected No. of Points	Booklet	Expected No. of Points
	411		30.5		
	410				32.0
	409		30.0		
	408				
	407		29.5		31.5
	406				
	405	7C	29.0	7A	31.0
	404				
	403		28.5		30.5
	402		28.0		
High	401				30.0
	400		27.5		
	399				29.5
	398		27.0		
	397				29.0
	396		26.5		
	395				28.5
	394		26.0		
	393		25.5		28.0
	392				
	391		25.0		27.5
	390				
	389		24.5		27.0
	388				
	387		24.0		26.5
	386				26.0
	385		23.5		
	384		23.0		25.5
	383				
	382		22.5		25.0
	381				
	380		22.0		24.5
	379				
	378		21.5		24.0
	377				
	376		21.0		23.5
	375				
	374		20.5		23.0
	373				
	372		20.0		22.5
	371				
	370		19.5		22.0
	369				
	368	6C	19.0		21.5
	367				
Median -->	366		18.5	5A, 6A	21.0
	365				
	364	5C	18.0		20.5
	363				
	362				20.0
	361		17.5		
	360				
	359		17.0		19.5
	358				
	357		16.5		19.0
	356				
	355				18.5
	354		16.0		
	353				
	352		15.5		18.0
	351				
	350				17.5
	349	4C	15.0		
	348				
	347		14.5	4A	17.0
	346				
	345				16.5
	344		14.0		
	343				
	342				16.0
	341		13.5		
	340				
	339				15.5
	338		13.0		
	337				
	336				15.0
	335		12.5		
	334				
	333				14.5
	332	3C	12.0		
	331				
	330			3A	14.0
	329				
	328		11.5		
	327				13.5
	326				
Low	325				
	324	2C	11.0		
	323			2A	13.0
	322				
	321				
	320		10.5		12.5
	319				
	318				
	317				
	316		10.0		12.0
	315				

Figure 19: Proficient Booklet Score Chart for Group A showing the median (yellow highlight), high, and low Proficient cut scores (horizontal lines) and the location of a panelist's round 1 cut score (circled)

For each test form, the IST provided the score a student received (0 = incorrect, 1 = correct) for every score point on each student booklet. The items and score points were ordered from easiest to hardest, bottom to top, and the student booklets were ordered from lowest to highest scoring, left to right. Figure 20 illustrates the IST for Form C, the common form. Panelists could use the IST to see, at a glance, the response patterns of students across the range of the achievement scale. For example, in Figure 20 panelists could see that in one of the borderline Proficient booklets, booklet 6C, the student received credit for 28 of the 44 total points and correctly answered many of the easy items and fewer of the hard items.

In the whole-group exercise, the panelists reviewed the BSCs and ISTs in relation to the two student booklets at the Proficient cut score on the common form (booklets 5C and 6C in Figures 19 and 20). Using the IST, panelists were told to observe the response patterns of the two student booklets near the Proficient cut score (5C and 6C) and to note that:

- The students answered different items correctly and incorrectly, but the overall proportion of items answered correctly was nearly the same.
- Differences in correct and incorrect answers may be due to variation in student mastery across content areas.
- Students did not get all items below the round 1 Proficient cut score of 366 correct and all above incorrect, but the probability of a correct response increased the farther below the cut score an item was and decreased the farther above the cut score an item was.

Once they were able to understand and interpret the information provided in the IST, panelists were given the opportunity to independently review booklets 5C and 6C. They were instructed to take note of where their cut scores fell in relation to the scores on these booklets, and to consider if performance represented by the booklets was too high, too low, or just right for the lower borderline of Proficient. A brief discussion was held following this review, in which panelists shared their perceptions of the level of performance exhibited in the booklets as related to the performance described in the Proficient ALD. The purpose of the discussion was to help panelists begin the process of gaining a shared understanding of the meaning of borderline performance for the Proficient achievement level.

Following this discussion, panelists began an independent booklet review of all 20 booklets provided to their group. They were told to review at least two booklets at the borderline of each achievement level and one booklet in the middle of each level, including Below Basic. During this review, they were to consider:

- How performance at the group cut score differed from performance at the middle of an achievement level.
- How students at their round 1 cut score were performing in relation to students at the group cut score.
- If performance at the group cut score was higher, lower, or just right for the lower borderline of the achievement level, given the ALD.

2009 NAEP Science ALS
Item Score Table
Grade 4 – Form C

					Basic Cut			Proficient Cut			Advanced Cut		
	Scale			1C	2C	3C	4C	5C	6C	7C	8C	9C	10C
Handle	Value	Block	Seq	17	20	21	24	27	28	32	34	35	38
C21_3	476	2	9	0	1	0	0	0	1	1	0	1	0
C15_3	436	4	9	0	0	0	0	0	0	1	0	1	1
C13_2	435	4	4	0	0	0	0	0	0	0	0	1	1
M149	434	4	18	0	0	0	0	0	1	0	0	1	1
M148	431	2	17	0	0	0	1	0	1	1	1	1	1
C12_2	427	4	13	0	0	0	0	1	0	0	1	0	0
M143	419	4	3	0	1	1	1	1	0	0	1	0	0
C21_2	414	2	9	0	1	1	0	0	1	1	1	1	1
M134	408	4	12	0	0	0	0	0	0	1	1	0	0
M131	403	2	16	0	0	0	0	0	0	1	1	0	1
M132	403	4	17	0	0	1	0	0	0	1	1	1	1
M124	398	4	16	1	1	0	1	0	0	0	1	0	1
M118	396	4	7	1	0	0	0	0	1	0	1	0	1
M112	393	2	2	0	0	1	1	1	1	0	0	0	0
M113	393	2	15	0	1	0	1	1	1	1	0	1	1
M110	392	4	15	0	0	0	0	1	1	1	1	0	1
M106	391	4	14	0	0	1	1	1	1	1	1	1	1
M104	390	4	2	1	1	0	0	1	0	1	0	1	1
M101	389	2	10	1	0	0	1	1	0	1	1	1	1
M96	387	2	14	0	1	1	0	1	0	1	0	1	1
M93	386	2	8	1	0	0	1	0	0	0	0	1	0
C12_1	385	4	13	0	0	0	1	1	0	1	1	1	1
C15_2	383	4	9	0	0	1	1	0	0	1	1	1	1
C4_2	381	2	13	0	0	1	0	0	1	0	1	1	1
M81	378	4	1	1	1	1	0	0	1	1	1	1	1
M80	377	4	5	1	1	0	1	1	1	1	0	1	1
M67	368	2	11	1	0	0	0	1	0	0	1	1	1
M58	365	2	6	0	1	1	1	0	0	1	1	1	1
M60	365	4	8	0	1	0	1	1	1	1	1	1	1
M55	364	2	7	1	1	1	1	1	1	0	1	1	1
C21_1	360	2	9	1	1	1	0	0	1	1	1	1	1
M50	358	2	5	0	1	0	1	1	1	1	1	1	1
M42	353	2	12	1	0	0	1	1	1	1	1	1	1
M25	342	4	11	0	0	1	0	1	1	0	1	1	1
M20	339	2	18	0	0	1	0	1	1	1	1	1	1
C1_2	338	2	4	1	1	0	1	1	1	1	1	0	1
C15_1	325	4	9	1	0	1	1	1	1	1	1	1	1
C4_1	324	2	13	0	1	1	1	1	1	1	1	1	1
M7	320	2	1	1	1	1	0	1	1	1	1	1	1
M6	317	2	3	0	1	1	1	1	1	1	1	1	1
M5	313	4	6	1	0	1	1	1	1	1	1	1	1
C13_1	312	4	4	1	1	1	1	1	1	1	1	1	1
C1_1	290	2	4	1	1	0	1	1	1	1	1	1	1
M1	285	4	10	0	0	1	1	1	1	1	1	1	1

Figure 20: Item Score Table for form C, with the items listed from hardest to easiest (top to bottom) in the left most column and the student booklets listed from lowest to highest scoring (left to right) in the top row with the number correct score on the booklet provided below the booklet identifier (1C to 10C)

At the conclusion of the independent review, panelists discussed with each other these questions and shared their reactions to the performance exhibited in the booklets. They were told that their task was to share their thoughts, but not to convince one another, and that the purpose of the discussion was to give each of them further information and insight to incorporate into their round 2 cut score recommendations.

Round 2 Cut Score Recommendations

In making round 2 cut score recommendations, panelists were instructed to work independently. Beginning with Proficient, then Basic, then Advanced, panelists chose a scale value and recorded the scale value on their Cut Score Recommendation form. Panelists were instructed to circle the scale value they chose for their round 2 cut score recommendation on their BSC and Primary Item Map and to move their round 1 bookmark in their OIB to the last item in their OIB with the scale value less than or equal to their recommended cut score.

Specific instructions were provided to aid them in the selection of their round 2 cut scores. They were instructed to select a range of scale scores within which they were deliberating, their range of uncertainty. This range might encompass, for example, the panelists' own cut score at the low end and a booklet that they felt represented borderline performance at the high end. Once they had identified the range, they were to locate the high and low points of this range in their OIB and BSCs and to consider: (a) what a student needed to know and be able to do to answer correctly items at-or-below potential cut scores in the OIB, and (b) the performance associated with potential cut scores in the booklets indicated on the BSC.

In considering booklets, panelists were also reminded of a number of technical considerations. They were told that there are many different forms of the science assessment and each form has approximately 45 total points. Because the achievement scale represents a much larger range than 45 points, there are some achievement scale values for which there are not corresponding point values on the forms panelists are reviewing. These scale values may correspond to point values on different forms, however, and so panelists can, and should, consider interpolating between raw score points on any given form when adjusting cut scores.

Round 3: Consequences Data

Overview of Round 3

The third round of the ALS method started with presentation of the cut score results from round 2. Panelists then received holistic feedback in the form of consequences data, or the proportion of students performing at or above and within each achievement level, to help with their understanding of what students in the levels that they have now defined can do. Following a grade-group discussion of the consequences data, panelists selected a scale value representing their final cut score for each achievement level.

Feedback from Round 2

Feedback from round 2 was presented using the same materials and formats that were used to present feedback after round 1. Feedback from round 2 consisted of the median cut scores, the cut score distribution, and the location of their cut scores relative to the median cut scores. Panelists were given a new Primary Item Map, Booklet Score Charts, and Booklet Score Plots on which the round 2 cut scores were marked. A table of the group cut scores from rounds 1 and 2 was presented to show panelists the current group cut scores and how the cut scores had changed over rounds. Panelists then identified where their cut scores fell in relation to the group cut scores on the Primary Item Map, Booklet Score Charts, and OIB.

Consequences Data and Discussion

The percent of students within and at or above each achievement level, including Below Basic, were reported to panelists as consequences data. These percentages were based on the 2009 distribution of student performance relative to the round 2 group cut scores.[1] The consequences data were presented to panelists in round 3 in the format shown in Figure 21. Panelists were also instructed to write the percentages of students in each achievement level and Below Basic in the left margin of their Primary Item Map.

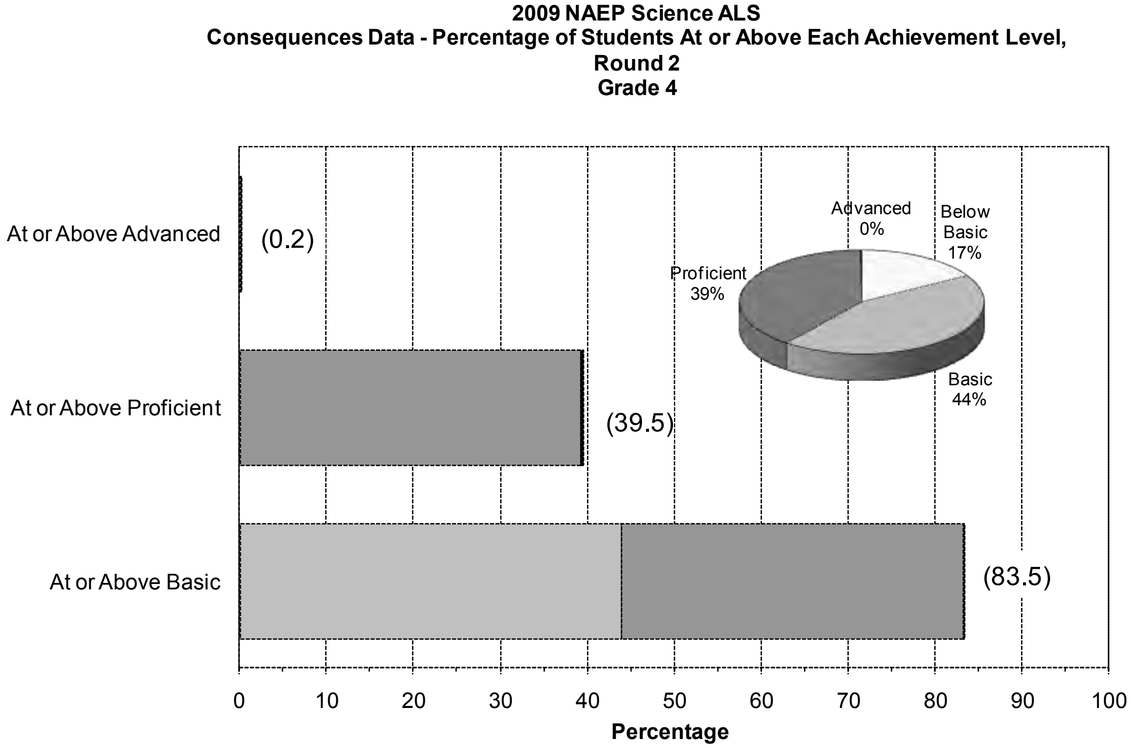


Figure 21: Example of consequences data presented to panelists in round 3

The consequences data were discussed prior to panelists making their round 3 cut score recommendations. As a lead-in to the discussion, panelists were told that the data came from the 2009 administration of the science NAEP. The sample was nationally representative, and panelists were told to keep in mind that student performance was influenced by student motivation and by the amount of time available. But regardless of what students can do as illustrated by the consequences data, it's what students should be able to do, according to the ALDs that rule the day. The discussion was largely left open to panelists, but a number of questions were suggested for discussion. These included: How do you feel about these cut scores now that you have seen the consequences data? Are you surprised by the percentages? Are these consequences about what you expected for a nationally representative sample of students? How are your expectations influenced by your own experiences? What allowance, if any, should be made for motivation? For the timed conditions of the test?

[1] Due to miscommunication of the 2009 NAEP reporting scale transformations, there were small differences between the percentages reported to the panelists and the actual percentages based on the 2009 distribution of student performance. See *Developing Achievement Levels on the 2009 National Assessment of Educational Progress in Science for Grades Four, Eight, and Twelve: Technical Report* (ACT, 2010) for details about these differences. There is evidence that these small differences did not affect panelists' interpretation of the cut scores they selected.

Round 3 Cut Score Recommendations

The purpose of round 3 cut score recommendations was to allow panelists to adjust their cut score recommendations based on feedback after round 2, including the consequences data. Panelists were instructed to work independently, study the feedback from round 2, reflect on the discussion of the consequences data, and determine if they felt their round 2 cut score recommendations needed to be changed. If they chose to change any of their recommendations, they were instructed to consult their OIB and item map to determine if the new cut scores they were considering were consistent with performance described in the ALDs. Panelists then recorded their cut score recommendations as they did in round 2.

Post-Round 3 Activities

After round 3, panelists were provided with their final cut scores, the cut score distribution, and the consequences data associated with the final cut scores. Next, they were asked to complete a Consequences Questionnaire, which is designed to elicit panelists' opinions on the suitability of the cut scores and whether they would choose a different cut score. Finally, panelists were asked to rate potential exemplar items.

Feedback from Round 3

Feedback from round 3 was given in the usual fashion except that panelists did not complete rater location tasks, identifying where their cut scores fell in relation to the final group cut. Panelists were given a new Primary Item Map with the final cut scores derived from round 3 and a new Cut Score Distribution Chart. They were instructed to remove their bookmarks from their OIB and to discard those bookmarks. They were then told to move the group bookmarks to the final cut scores. This was to emphasize that the round 3 cut scores were the final cuts. The feedback also included consequences data based on the round 3 group cut scores. This was presented in the format shown in Figure 21.

Panelists were told that the round 3 group cut scores would be reported to the Governing Board as one of the key outcomes of the ALS meeting. It was very important that panelists understood the level of performance exhibited by students at the cut scores, which was the purpose of the feedback, and that they evaluated the cut scores based on the match between the criterion-referenced feedback, the ALDs, and their concept of borderline performance.

Consequences Questionnaire

The purpose of the consequences questionnaire was to provide the Governing Board with information about panelists' reactions to the final consequences data. A copy of the Consequences Questionnaire is included in Appendix K. The cut scores and the achievement level percentages were filled in for the panelists. The questionnaire asked panelists if they would want to make changes to any of the cut scores after learning the consequences of their cut scores. Panelists could recommend a different cut score to represent each achievement level for any or all three cut scores. A Cut Score Proportion Chart (Figure 22) was provided to allow panelists to see the relative impact of changing from one cut score to another if they wanted to raise or lower the cut score. This chart provided the percentage of students scoring at or above every fifth score value on the ACT NAEP-like assessment scale. The final cut scores were marked on the chart. Panelists were instructed to use this information to help them decide what final cut scores they would recommend after having learned the achievement level percentages associated with their round 3 cut scores.

2009 NAEP Science ALS
Consequences Proportions Chart, Round 3
Grade 4

	NAEP-like Score	Percent at or Above
Advanced	510	0.0
	505	0.0
	500	0.0
	495	0.0
	490	0.0
	485	0.0
	480	0.0
	475	0.0
	470	0.0
	465	0.0
	460	0.0
	455	0.0
	450	0.0
Proficient	445	0.0
	440	0.0
	435	0.0
	430	1.0
	425	2.0
	420	3.0
	415	5.0
	410	7.0
	405	11.0
	400	14.0
	395	19.0
	390	24.0
	385	29.0
	380	34.0
Basic	375	40.0
	370	46.0
	365	52.0
	360	57.0
	355	62.0
	350	67.0
	345	72.0
	340	76.0
	335	80.0
	330	83.0
	325	86.0
	320	88.0
	315	91.0
	310	92.0
	305	94.0
	300	95.0
	295	96.0
	290	97.0
	285	97.0
	280	98.0
	275	98.0
	270	99.0
	265	99.0
	260	99.0
	255	99.0
	250	99.0
	245	99.0
	240	99.0
	235	99.0
	230	99.0
	225	99.0
	220	100.0
	215	100.0

Figure 22: Illustration of a Cut Score Proportion Chart identifying the percent of students scoring at or above every fifth score level [2]

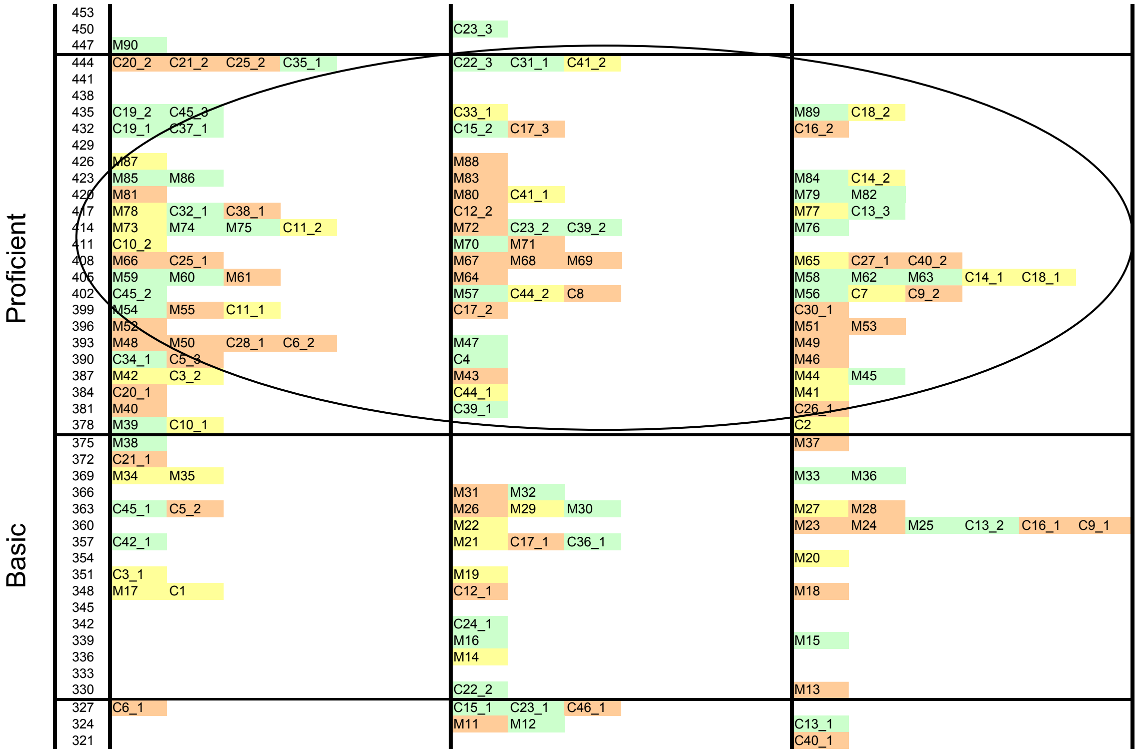
Ratings of Exemplar Items

The purpose of the exemplar item rating task was to provide the Governing Board with information concerning the suitability of items for illustrating what students in the achievement levels know and can do. The panelists had spent many hours working with the items and the

[2] See the *Technical Report* (ACT, 2010) for details about the differences in the Cut Score Proportion Charts provided to the panelists and charts based on the 2009 distribution of student performance.

ALDs, translating their meaning into cut scores. They were in a good position to provide the Governing Board with this input.

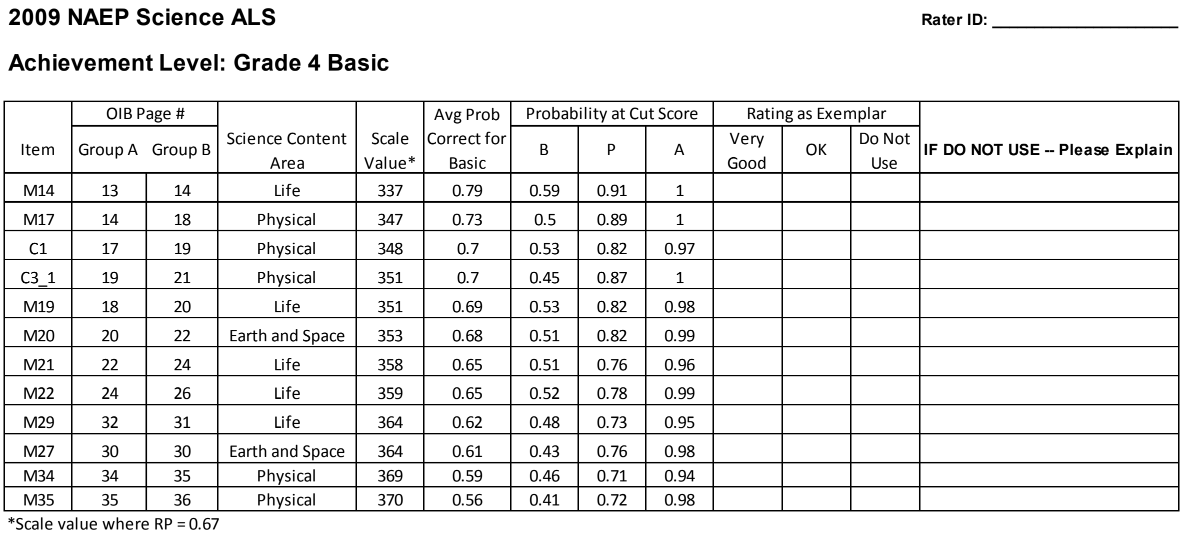
Potential exemplars for each grade were drawn from two blocks of the assessment that were selected for possible release to the public. COSDAM specified that potential exemplar items would be identified only from within the achievement level, such that a student at the top of that level would have at least a 67% chance of answering the item correctly. That is, an item was selected as a potential exemplar for an achievement level if it mapped to that achievement level and not to a lower or higher level (see Figure 23). This criterion produced reasonable-sized pools of items for potential use as exemplars.



	453			
	450		C23_3	
	447	M90		
Proficient	444	C20_2 C21_2 C25_2 C35_1	C22_3 C31_1 C41_2	
	441			
	438			
	435	C19_2 C45_3	C33_1	M89 C18_2
	432	C19_1 C37_1	C15_2 C17_3	C16_2
	429			
	426	M87	M88	
	423	M85 M86	M83	M84 C14_2
	420	M81	M80 C41_1	M79 M82
	417	M78 C32_1 C38_1	C12_2	M77 C13_3
	414	M73 M74 M75 C11_2	M72 C23_2 C39_2	M76
	411	C10_2	M70 M71	
	408	M66 C25_1	M67 M68 M69	M65 C27_1 C40_2
	405	M59 M60 M61	M64	M58 M62 M63 C14_1 C18_1
	402	C45_2	M57 C44_2 C8	M56 C7 C9_2
	399	M54 M55 C11_1	C17_2	C30_1
	396	M52		M51 M53
	393	M48 M50 C28_1 C6_2	M47	M49
	390	C34_1 C5_3	C4	M46
	387	M42 C3_2	M43	M44 M45
	384	C20_1	C44_1	M41
	381	M40	C39_1	C26_1
	378	M39 C10_1		C2
Basic	375	M38		M37
	372	C21_1		
	369	M34 M35		M33 M36
	366		M31 M32	
	363	C45_1 C5_2	M26 M29 M30	M27 M28
	360		M22	M23 M24 M25 C13_2 C16_1 C9_1
	357	C42_1	M21 C17_1 C36_1	
	354			M20
	351	C3_1	M19	
	348	M17 C1	C12_1	M18
	345			
	342		C24_1	
	339		M16	M15
	336		M14	
	333			
	330		C22_2	M13
	327	C6_1	C15_1 C23_1 C46_1	
	324		M11 M12	C13_1
	321			C40_1

Figure 23: Illustration of the range used for selection of potential exemplar items for the Proficient level

Figure 24 shows an example of the Exemplar Item Rating form panelists were given for rating items. Panelists were given separate forms for the items associated with each achievement level. The form listed the items/score points in the order they appeared in the OIB, and identified the items by handle and the OIB page number where they could be found. The form provides the science content area the item is measuring, the scale value to which the item maps, the average probability that students within the achievement level have of getting the item correct, and the probability that students scoring at each cut score have of getting the item correct.



2009 NAEP Science ALS Rater ID: ___________________

Achievement Level: Grade 4 Basic

Item	OIB Page #		Science Content Area	Scale Value*	Avg Prob Correct for Basic	Probability at Cut Score			Rating as Exemplar			IF DO NOT USE -- Please Explain
	Group A	Group B				B	P	A	Very Good	OK	Do Not Use	
M14	13	14	Life	337	0.79	0.59	0.91	1				
M17	14	18	Physical	347	0.73	0.5	0.89	1				
C1	17	19	Physical	348	0.7	0.53	0.82	0.97				
C3_1	19	21	Physical	351	0.7	0.45	0.87	1				
M19	18	20	Life	351	0.69	0.53	0.82	0.98				
M20	20	22	Earth and Space	353	0.68	0.51	0.82	0.99				
M21	22	24	Life	358	0.65	0.51	0.76	0.96				
M22	24	26	Life	359	0.65	0.52	0.78	0.99				
M29	32	31	Life	364	0.62	0.48	0.73	0.95				
M27	30	30	Earth and Space	364	0.61	0.43	0.76	0.98				
M34	34	35	Physical	369	0.59	0.46	0.71	0.94				
M35	35	36	Physical	370	0.56	0.41	0.72	0.98				

*Scale value where RP = 0.67

Figure 24: Illustration of Exemplar Item Rating form

Panelists were instructed to discuss each potential exemplar item with their table group. Did the science content and practices required by the item seem appropriately matched to the ALD for the achievement level? Following this discussion, panelists provided independent ratings on their Exemplar Item Rating form as to whether or not they considered the item or score point to be a good exemplar for that achievement level. Panelists marked *Very Good* or *OK* for items that they would recommend as exemplars and *Do Not Use* for items that they would not recommend as exemplars.

Process Evaluations

The validity of standard-setting outcomes depends in part on what is called *procedural validity*. Procedural validity is provided in the form of evidence that the procedures were carried out as intended and were understood by the panelists. At the end of each round and each day, panelists were provided with an evaluation form designed to assess their understanding of instructions, tasks, and materials. There were a total of five questionnaires administered over the course of the meeting. Most responses were collected on Likert scales, but several responses were narratives that addressed specific aspects of the process. These evaluations were reviewed at the end of each day and any sources of confusion were identified for clarification with individual panelists or the group as a whole. The five process evaluation questionnaires are presented in their entirety in Appendix L. The results for each questionnaire are also included in Appendix L for each grade. For each question, the appendix shows the frequency of responses per Likert-scale category and the average response.

Evaluation of the Method Outcomes

In order to allow for comparison of procedural data from the 2009 science ALS with methods used in previous NAEP standard-setting meetings, an effort was made to ensure that a number of the evaluation questions were largely the same as questions used to evaluate NAEP ALS methods in the past. Strong support for procedural validity would be demonstrated by consistent mean (average) responses on most items at or above 4.0 on a 1–5 scale. In general, panelist evaluations of the ALS were comparable to or better than the evaluations from the 1996 science, the 2005 grade 12 mathematics, and the 2006 grade 12 economics standard setting contracts. Based on these results, it seems reasonable to conclude that, in panelists' perceptions, the quality of the ALS process for science equals the quality of the methods used to establish cut scores for the NAEP in other subjects.

The Mapmark ALS process compared well with methods ACT used in past standard-setting work for the Governing Board. Key evaluation questions on the last process evaluation questionnaire addressed panelists' overall perception of the effectiveness of the ALS method, whether the process afforded them the opportunity to use their best judgment, whether the process yielded defensible and reasonable cut scores that represented meaningful distinctions between achievement levels, and whether they were confident in their final cut scores. Responses were on a Likert scale of 1–5, with 5 the highest level of agreement.

Figure 25 shows the average ratings for grade 12 from the 2009 science ALS and some previous ALS studies on these key overall process evaluation questions. The method used in the 2006 grade 12 economics ALS process was Mapmark with Whole Booklet Feedback, the method used in the 2005 grade 12 mathematics ALS process was Mapmark with Domains, and the method used in the 1996 science ALS process was a modified-Angoff method. Tests of statistical significance were not performed on the differences among methods, but it can be seen that the average rating for the Mapmark with Whole Booklet Feedback method used in the 2009 ALS compared favorably with the averages for the other ALS processes.

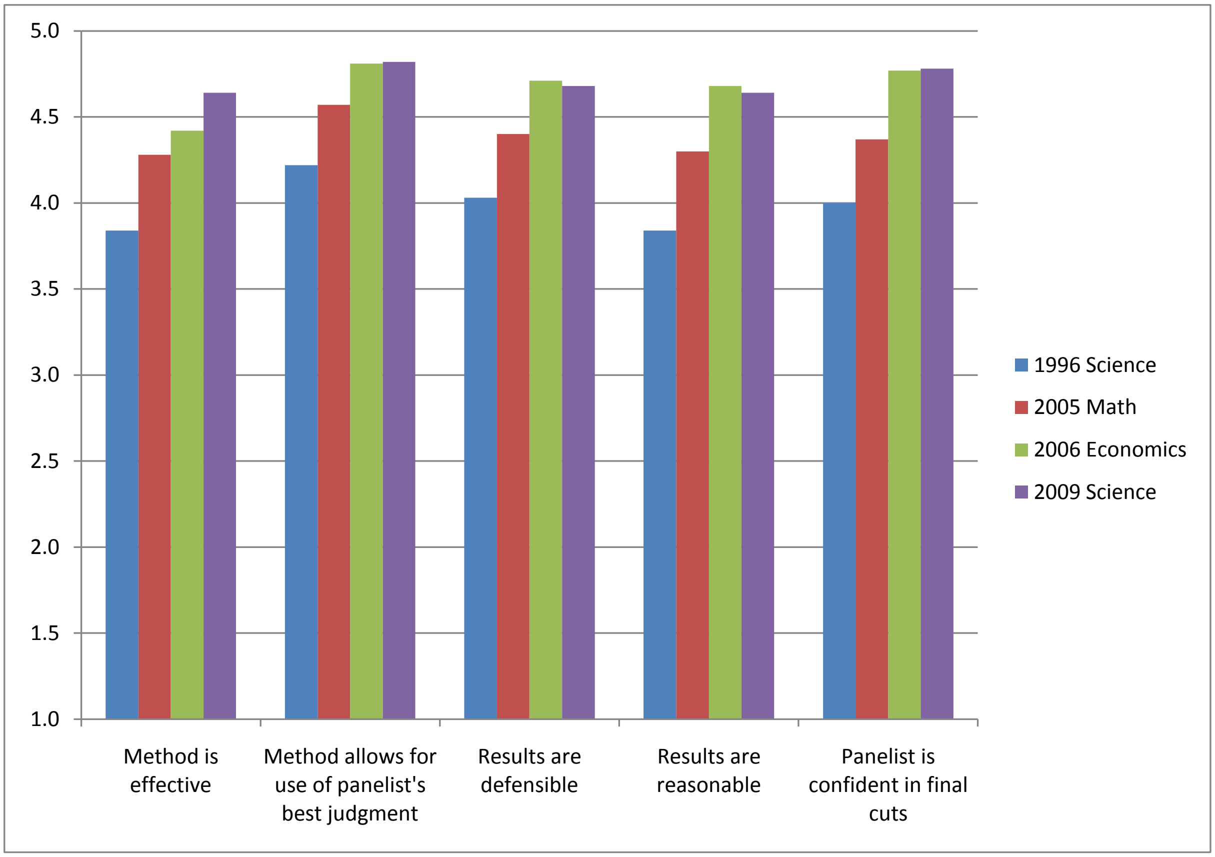


Figure 25: Average ratings for grade 12 from 2009 science ALS and previous ALS studies on key process outcome questions

In addition, most panelists said they would be willing to sign a statement recommending the use of the achievement levels resulting from the standard-setting procedure. Responses were on a Likert 1-4 scale, with 4 being the most positive response. Table 18 shows responses and mean ratings on the endorsement question from the 2009 science ALS and the same previous ALS contracts. The rates of endorsement (93% to 100% favorable) compare well with previous standard-setting processes that ACT has conducted for the Governing Board.

Table 18: Percentages of panelists endorsing group cut scores for various ALS studies

	I would be willing to sign a statement (after reading it of course) recommending the use of the cut scores resulting for this ALS process.				
ALS Study	**Yes, definitely (4)**	**(3)**	**(2)**	**No, definitely not (1)**	**Average Rating**
2009 Science Grade 4	66	28	7	0	3.59
2009 Science Grade 8	70	30	0	0	3.70
2009 Science Grade 12	89	11	0	0	3.89
2006 Economic Grade 12	77	20	3	0	3.73
2005 Mathematics Grade 12	66	31	3	0	3.62
1996 Science Grade 4	41	52	3	0	3.28
1996 Science Grade 8	52	42	3	0	3.39
1996 Science Grade 12	53	38	6	3	3.41

Clarity of Instructions and Presentations

Table 19 shows average ratings on process evaluation questions pertaining to clarity of presentations on certain topics addressed during the orientation sessions the first day of the ALS meeting. The presentations were consistently rated as clear. The panelists from all three grades attended the same orientation sessions so any variation in the ratings across grade levels reflects differences across the three groups of panelists, not differences in presenters or presentations.

Table 19: Average ratings of clarity of topic presentation
The explanation/overview/presentation of the . . . was
(5 = *absolutely clear*; 3 = *somewhat clear*; 1 = *not at all clear*)

Round	Question #	Orientation Topic	ALS Average Rating		
			Grade 4	Grade 8	Grade 12
Pre	1-4	NAEP in general	4.67	4.48	4.18
Pre	1-5	Development of the Science NAEP	4.53	4.44	4.38
Pre	1-6	Major organizations involved and the roles of each	4.50	4.44	4.46
Pre	1-15	Overview of method to be followed in the meeting	4.03	3.69	3.93
Pre	1-16	How an item map is constructed	4.11	3.96	3.86
Pre	1-18	Information in the OIB	4.45	4.04	4.17
Pre	1-20	Science Framework	4.52	4.15	4.28

Table 20 shows average ratings on process evaluation questions pertaining to clarity of instructions for the major tasks completed during the ALS process. In general, all of the ratings indicate that the panelists found the instructions clear as they went through the ALS process.

Table 20: Average ratings of clarity of instructions by task
The instructions on (what/how I/we was/were to do in/for the . . .) were
(5 = *absolutely clear*; 3 = *somewhat clear*; 1 = *not at all clear*)

Round	Question #	Task	ALS Average Rating		
			Grade 4	Grade 8	Grade 12
1	1-23	Grade-group item review	4.03	3.07	3.93
1	2-5	Independent OIB review	4.47	3.93	4.50
1	2-11	Table discussion of OIB	4.33	3.59	4.36
1	2-31	Placing the bookmarks	3.45	4.00	3.82
2	3-6	Borderline Proficient exercise	4.00	4.22	4.46
2	3-13	Table-group whole booklet review	4.17	4.11	4.59
2	3-27	Recommending round 2 cut scores	4.34	4.67	4.73
3	4-8	Using the consequences data	4.53	3.74	4.64
3	4-16	Recommending the final cut scores	4.76	4.59	4.82
Post	5-5	Completing consequences questionnaire	4.90	4.78	4.46
Post	5-7	During each round	4.50	4.19	4.64
Post	5-17	Exemplar item rating task	4.70	4.37	4.82

At the conclusion of the ALS process, panelists were also asked to rate instructions and their understanding of tasks for the entire process. They were asked to indicate the degree to which they felt the instructions on what they were to do during each round were clear (1 = *not at all clear* to 5 = *absolutely clear*) and the adequacy of their understanding of the tasks they were to accomplish during each round (1 = *totally inadequate* to 5 = *totally adequate*). Results for these two evaluation questions are shown in Figure 26. Overall, panelists for all three grades indicated that they found the instructions clear and understood the tasks they were to accomplish during each round of the ALS process.

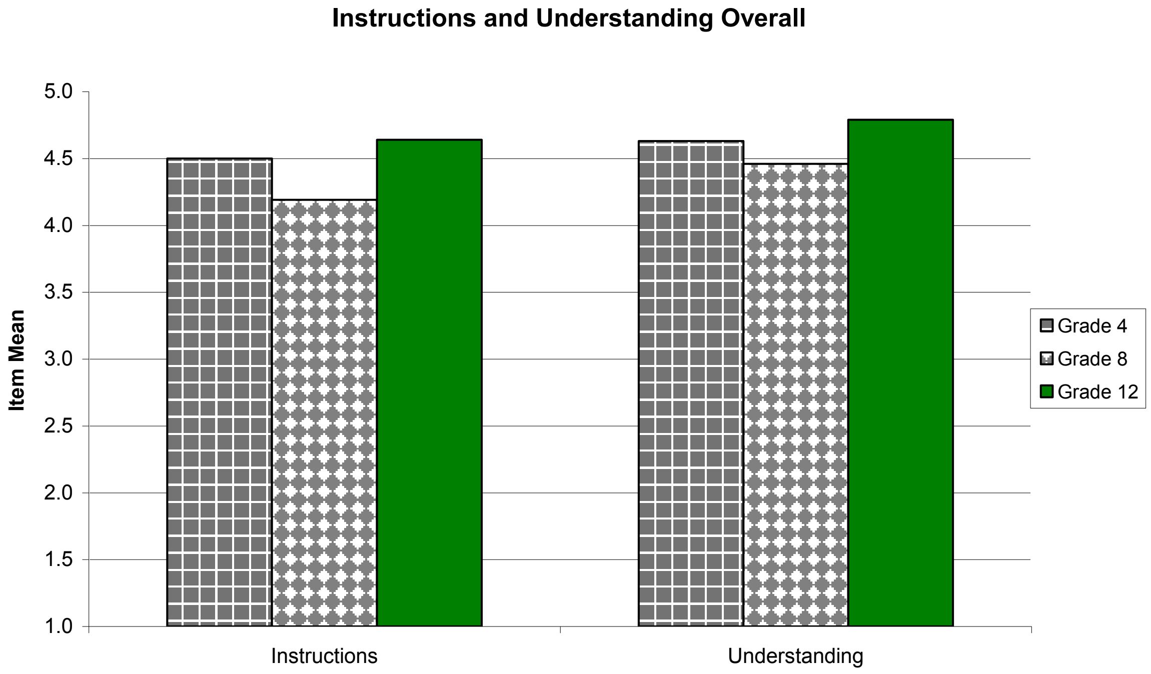


Figure 26: ALS average ratings on clarity of instructions and panelist understanding of tasks

Understanding of Concepts and Feedback

Understanding of concepts and feedback depends on the clarity of presentations and instructions, which the previous section shows was good. It can be seen in Table 21 that panelists had a good understanding of concepts in the ALS process. In particular, understanding of concepts unique to the Mapmark process—such as the concept of how to use item maps and of the information in Booklet Score Charts, Booklet Score Plots, and Item Score Tables—was high, as indicated by average ratings above 4.5 in Table 21.

Table 21: Average ratings of understanding of concepts
I understand/understood . . .
(5 = *totally agree*; 3 = *somewhat agree*; 1 = *totally disagree*)

Round	Question #	Concept	ALS Average Rating		
			Grade 4	**Grade 8**	**Grade 12**
1	1-10	Difference between criterion-referenced and norm-referenced standards	4.55	4.35	4.34
1	2-3	Score levels of constructed response items	3.83	4.00	3.93
2	2-6	How to use my item map with the OIB	4.83	4.56	4.71
2	2-32	How to use the ALDs to choose my bookmarks	4.27	4.19	4.37
3	3-19	Information in the Booklet Score Chart	4.55	4.63	4.74
3	3-20	Information in the Booklet Score Plots	4.59	4.74	4.81
3	3-21	Information in Item Score Tables	4.66	4.74	4.85
3	3-24	Difference between borderline performance and typical performance within an achievement level	4.62	4.52	4.70
Post	5-21	Purpose of this meeting	4.93	4.96	4.89

Panelists had good understanding of the feedback they were given. As shown in Table 22, average ratings of understanding of general types of feedback such as the group cut scores (round __ median cut scores), rater location feedback, and consequences data were well above 4.0 after round 1 and tended to increase with each round.

Table 22: Average ratings of understanding of feedback
I understand/understood the round __ . . .
(5 = *totally agree*; 3 = *somewhat agree*; 1 = *totally disagree*)

Feedback	Grade	ALS Average Rating		
		Round 1	Round 2	Round 3
Median cut scores	4	4.73	4.93	4.93
	8	4.70	4.81	4.93
	12	4.82	5.00	5.00
What students at the round __ cut scores can do	4	4.50	4.80	4.87
	8	4.37	4.70	4.85
	12	4.68	4.79	4.89
Rater location feedback (where my round __ cut scores were in comparison to the median)	4	4.73	4.83	n/a
	8	4.74	4.81	n/a
	12	4.86	4.96	n/a
Cut score distribution chart	4	4.83	4.93	n/a
	8	4.89	4.93	n/a
	12	4.86	4.96	n/a
Consequences data	4	n/a	4.93	4.77
	8	n/a	4.81	4.78
	12	n/a	5.00	4.75

Understanding the ALDs and Borderline Performance

Panelist understanding of both the ALDs and the concept of performance at the lower borderline at each achievement level was also assessed. These are two concepts critical to the process of identifying cut scores. As expected, their understanding of these two critical concepts increased across rounds.

At the conclusion of round 1, panelists were asked to rate their understanding of the ALDs for each level (Basic, Proficient, and Advanced). Panelist responses were used to assess if clarification was needed during the meeting for any one level. For rounds 2 and 3, panelists were asked to rate their understanding of the ALDs for all levels combined. Table 23 shows that panelist understanding of the ALDs increased across rounds.

Table 23: Average ratings of understanding ALDs
My understanding of ALDs [in round __] was . . .
(5 = *totally adequate* to 1 = *totally inadequate*)

Grade	ALS Average Rating				
	Basic Round 1	Proficient Round 1	Advanced Round 1	Round 2	Round 3
4	4.03	4.07	4.03	4.66	4.80
8	4.11	4.26	4.19	4.59	4.70
12	4.41	4.37	4.33	4.63	4.89

Table 24 shows that the perceived consistency between the ALDs and panelists' cut score recommendations increased over rounds. These results are what one would expect from the patterns of understanding and concept formation evident in previous tables in this section.

Table 24: Average ratings of consistency of cut score recommendations with ALDs
I believe my round __ bookmark placements/cut score
recommendations are consistent with the ALDs
(5 = *totally agree,* 3 = *somewhat agree,* 1 = *totally disagree*)

	ALS Average Rating		
Grade	**Round 1**	**Round 2**	**Round 3**
4	3.83	4.55	4.73
8	3.74	4.78	4.74
12	4.14	4.59	4.82

At the conclusion of each round, panelists were also asked to respond to statements about performance at the lower borderline. The lower borderline question is worded slightly differently in the first round than it is in the second and third rounds. At the conclusion of the first round, panelists were asked, for each achievement level, to indicate their level of agreement (5 = *totally agree*, 1 = *totally disagree*) with: I was comfortable using the concept of performance at the lower borderline of _____. At the conclusion of the second and third rounds, the question was asked for all three levels combined. Panelists had to respond on a scale from 1 to 5 (1 = *not well formed*, 5 = *very well formed*) to the statement: At the time I provided my round ___ cut score recommendations, my concept of the lower borderline performance of an achievement level was ___. The mean panelist rating for these questions, by round, is provided in Table 25. The panelist ratings increase by round, as is consistent with patterns of response seen in previous standard setting meetings.

Table 25: Average ratings of development of concept of borderline performance

	ALS Average Rating				
Grade	**Basic Round 1**	**Proficient Round 1**	**Advanced Round 1**	**Round 2**	**Round 3**
4	3.83	3.87	3.83	4.48	4.63
8	3.89	4.00	3.89	4.52	4.67
12	4.11	4.32	4.11	4.59	4.82

Comfort and Confidence

As shown in Table 26, panelists were comfortable with key features of the Mapmark process including the value of the RP criterion (0.67) and its meaning (mastery). In addition, panelists' confidence in their cut score recommendations (see Table 27) increased steadily from round 1 to round 3. The trend of increasing confidence in cut scores over rounds is typical of other methods and standard-setting meetings ACT has conducted for the Governing Board.

Table 26: Average ratings of comfort level with various features of Mapmark
I think I will be/I was comfortable . . .
(5 = *totally agree*; 3 = *somewhat agree*; 1 = *totally disagree*)

Round	Question #	Task	ALS Average Rating		
			Grade 4	Grade 8	Grade 12
1	1-17	Using a 2/3 or 0.67 probability to interpret the location of an item on my map	4.07	4.04	4.03
1	2-7	Working through the OIB on my own	4.73	4.59	4.85
1	2-38	Using a 0.67 probability to define mastery in placing my bookmark	3.87	3.56	4.04
2	3-31	Choosing scale values instead of placing bookmarks to recommend cut scores	4.57	4.58	4.64

Table 27: Average ratings of confidence level in cut scores by round

Grade	ALS Average Rating		
	Round 1	Round 2	Round 3
4	3.33	4.20	4.70
8	3.37	4.31	4.56
12	3.68	4.64	4.93

Usefulness/Helpfulness of Activities and Information

Results in Table 28 show that panelists found the grade-group and table-group item review activities to be useful. During these activities panelists worked together to identify the science content and practices that a student needs to know in order to answer each item correctly. Table 29 shows that panelists generally found the information and materials in the Mapmark process to be helpful. Average ratings for all materials and information specific to the Mapmark process were above 4.0 and much higher than the average rating for the helpfulness of consequences data (the percent of students in achievement levels). This may be regarded as a positive outcome since the consequences data are purely normative information. As in previous Mapmark standard settings, the OIB was perceived to be most helpful, followed closely by the ALDs and the Primary Item Map, except for grade 8 where the ALDs were perceived to be most helpful.

Table 28: Average ratings of usefulness of activities

The ________ was

(5 = *very useful*; 3 = *somewhat useful*; 1 = *not at all useful*)

Round	Question #	Activity	ALS Average Rating		
			Grade 4	Grade 8	Grade 12
1	1-25	Grade-group work on constructed-response items	4.63	3.70	4.21
1	2-2	Table-group review of remaining constructed-response items	4.37	3.78	4.25
1	2-12	Table discussion of OIB	4.43	4.15	4.48

Table 29: Average ratings of helpfulness of information

During the ALS process, I found the ____________

(5 = *very helpful*; 3 = *somewhat helpful*; 1 = *not at all helpful*)

Round	Question #	Information	ALS Average Rating		
			Grade 4	Grade 8	Grade 12
Post	5-31	ALDs	4.87	4.93	4.75
Post	5-32	Ordered Item Book	4.93	4.89	4.82
Post	5-33	Primary Item Map	4.77	4.56	4.50
Post	5-34	Rater location data	4.70	4.41	4.21
Post	5-35	Consequences data	3.70	3.59	2.54
Post	5-36	Booklet Score Charts	4.43	4.15	4.36
Post	5-37	Booklet Score Plots	4.57	4.07	4.43
Post	5-38	Cut Score Distribution Chart	4.23	4.11	4.36

Independence of Judgment and Perspective

Process evaluation results indicated that the general instructions panelists were given with regard to maintaining their perspective and independent judgment were effective. As shown in Table 30, panelists tended to disagree with the statement that they felt pressure to recommend cut scores that were close to those of other panelists.

Table 30: Average ratings of perceived influences/pressure on cut score recommendations

I felt pressure to recommend bookmarks/cut scores that were
close to those recommended by other panelists
(5 = *totally agree*; 3 = *somewhat agree*; 1 = *totally disagree*)

		ALS Average Rating		
Round	**Question #**	**Grade 4**	**Grade 8**	**Grade 12**
1	2-34	1.20	1.17	1.14
2	3-30	1.17	1.12	1.43
3	4-19	1.13	1.30	1.39

At the conclusion of round 1, the average response to the question, I feel that my perspective is being heard by others in my table group (5 = *totally agree*), was 4.83, 4.56, and 4.89 for grades 4, 8, and 12, respectively. At the conclusion of the meeting, the average response to the statement, I felt my input was valued and considered by others in my group (5 = *to a great extent*), was 4.87, 4.67, and 4.75 for grades 4, 8, and 12, respectively.

Amount of Time Allocated for Tasks

Because of the large number of CR score points with very complex scoring rubrics and, consequently, Pilot Study timing issues, the adequacy of time allocated for tasks was an important issue for the ALS. The changes made to the ALS process and agenda did result in the panelists being able to complete the constructed-response item review task during the ALS meeting. This is reflected in the satisfactory results for the ALS for the process evaluation timing questions for the orientation and item review sessions. Details concerning the amount of time allocated for activities are presented in Table 31. Average ratings in this table indicate that the timing was appropriate for all activities as all but two average ratings fell between 2.5 and 3.5.

Table 31: Average ratings of amount of time allocated for activities
(5 = *far too long*; 3 = *about right*; 1 = *far too short*)

			ALS Average Rating		
Round	**Question #**	**Activity**	**Grade 4**	**Grade 8**	**Grade 12**
Pre	1-3	General orientation to NAEP	3.40	3.44	3.21
Pre	1-14	Mapmark method orientation	3.41	3.27	3.14
Pre	1-19	Science Framework presentation	3.38	3.38	3.21
Pre	1-22	Grade-group item review	2.97	2.88	2.76
1	2-1	Table-group item review	2.67	2.74	2.96
1	2-4	Independent OIB review	2.60	2.81	3.11
1	2-10	Table discussion of OIB	2.87	2.85	2.93
1	2-16	ALD presentation	3.47	3.85	3.15
1	2-30	Placing the bookmarks	2.87	2.67	3.14
2	3-5	Borderline proficient booklet exercise	3.34	3.26	3.32
2	3-12	Table-group whole booklet review	3.10	3.19	3.22
2	3-26	Round 2 cut score recommendations	3.21	3.33	3.33
3	4-9	Discussion of consequences data	3.13	2.19	3.14
Post	5-3	Consequences questionnaire	3.20	2.85	3.39
Post	5-10	Complete tasks during each round	2.90	3.04	3.21
Post	5-16	Exemplar item rating task	3.00	3.19	3.39

Reactions to Consequences Data

In the round 3 whole-group discussion of consequences data—the percent of students at or above each of the achievement levels—panelists did not express surprise at the very small percentage of students performing at the Advanced level. In addition, panelists questioned the appropriateness of getting this norm-referenced data during the process since the process is criterion-referenced, anchored to the ALDs. It was not surprising then that the group cut scores did not change much from round 2 to 3. The Basic cut score decreased by 2 points for grade 4; the Proficient cut score decreased by 1 point and the Advanced cut score decreased by 5 points for grade 8; and, there were no changes in cut scores from round 2 to 3 for grade 12. These results, along with comments voiced during the whole-group discussions, indicate that panelists were not unduly influenced by the introduction of consequences data in the process and maintained their commitment to criterion-referenced cut score judgments.

At the conclusion of the ALS, panelists were asked to review the achievement level percentages and to complete a questionnaire indicating the reasonableness of those percentages. They were asked to indicate if they felt that the percentages reflected their expectations about the proportions of students whose NAEP score would be at or above the cut score established for each achievement level and, if not, to indicate if they would raise or lower the cut scores to adjust the percentages. Results from the Consequences Questionnaire are summarized in Table 32. At least 70% of the panelists endorsed the final cut scores for grade 4; at least 93% of the panelists endorsed the final Basic and Proficient cut scores and 52% endorsed the Advanced cut score for grade 8; and at least 93% of the panelists endorsed the final cut scores for grade 12. Of those who recommended changes, several recommended changes for more than one level. In total, 18 (60%) grade 4 panelists, 13 (48%) grade 8 panelists, and 3 (11%) grade 12 panelists recommended changes. Most of the changes recommended were to make the final cut score the same as the panelist's round 3 cut score (77% grade 4, 57% grade 8, and 33% grade 12) suggesting that the consequences data had little impact on panelists' cut score recommendations.

Table 32: Panelists' responses on the ALS Consequences Questionnaire

Grade	Achievement Level	No. of Panelists Recommending Lower Cut Score		No. of Panelists Endorsing Final Cut Score with No Change	No. of Panelists Recommending Higher Cut Score	
		Lower by >10 Score Points	Lower by ≤10 Score Points		Raise by ≤ 10 Score Points	Raise by >10 Score Points
4	Basic	1	1	21	0	7
	Proficient	3	4	21	0	2
	Advanced	1	2	22	0	5
8	Basic	0	2	25	0	0
	Proficient	0	0	27	0	0
	Advanced	2	10	14	0	1
12	Basic	0	0	28	0	0
	Proficient	0	2	26	0	0
	Advanced	0	0	27	1	0

Three Research Questions

ACT was asked to investigate how the panelists understand and use the RP criterion as part of the process for setting achievement levels for the 2009 NAEP Science Assessment. Three questions were put forth as areas for inquiry:

1. How well do the standard-setting panelists understand and use the given RP criterion, and is using the criterion to set the standards more or less difficult than other tasks that may be required of the panelists?
2. Do panelists understand how the cut score might change if the RP criterion were changed?
3. Do the panelists understand how the knowledge, skills, and abilities (KSAs) for an item that falls within a particular achievement level relate to the ALDs of what students should know and be able to do at that level?

Information related to the three research questions was gathered via the process evaluation questionnaires administered during the ALS meeting and the Pilot Study. Questionnaires were administered at the end of each day (questionnaires 1, 2, 4, and 5) and/or at the end of each round (questionnaires 2, 3, and 4).

<u>Question 1</u>

To investigate Question 1, the panelists were asked for their evaluation of the difficulty of using some aspects of the standard setting that are thought to be more cognitively demanding. The specific aspects used were:

- Relating the science principles and practices for the individual items to the ALDs
- Using the ALDs for an achievement level to develop the idea of a minimally-qualified student
- Using the RP criterion to define mastery of a knowledge or skill
- Using the consequences data to adjust cut scores

For each of these four aspects, two distinct questions were asked in the process evaluation questionnaires. The first set of questions asked if the panelists were comfortable in using each aspect in the standard-setting process (response options ranged from *Totally Agree* to *Totally Disagree*). The second set of questions asked how difficult it was to take the aspect into account (response options ranged from *Not at All Difficult* to *Very Difficult*). These two questions were asked after each round in which the aspect was used. For the first three concepts, the questions appear in questionnaires 2, 3, and 4. The final aspect, using consequences data, was used only in round 3 (the last round where cut scores were chosen), and so, appears only in questionnaire 4.

To further address Question 1, the panelists were also asked to rank the aspects from most difficult to least difficult to use and to explain why the one that was considered most difficult was chosen. This question was asked after round 3 and appeared in questionnaire 4.

For Question 1, the ALS results showed that panelists were, in general, comfortable with the specified aspects of standard setting. The highest rating (easiest to use) was for the science principals and practices, and the lowest ratings (hardest to use) were for the consequences data. Figure 27 shows panelists' level of comfort with the concepts at the end of round 3 for all grades. For questions about difficulty, there were only minor differences across the four aspects considered. Figure 28 shows the ratings of difficulty in using the aspects for all grades at the end of round 3.

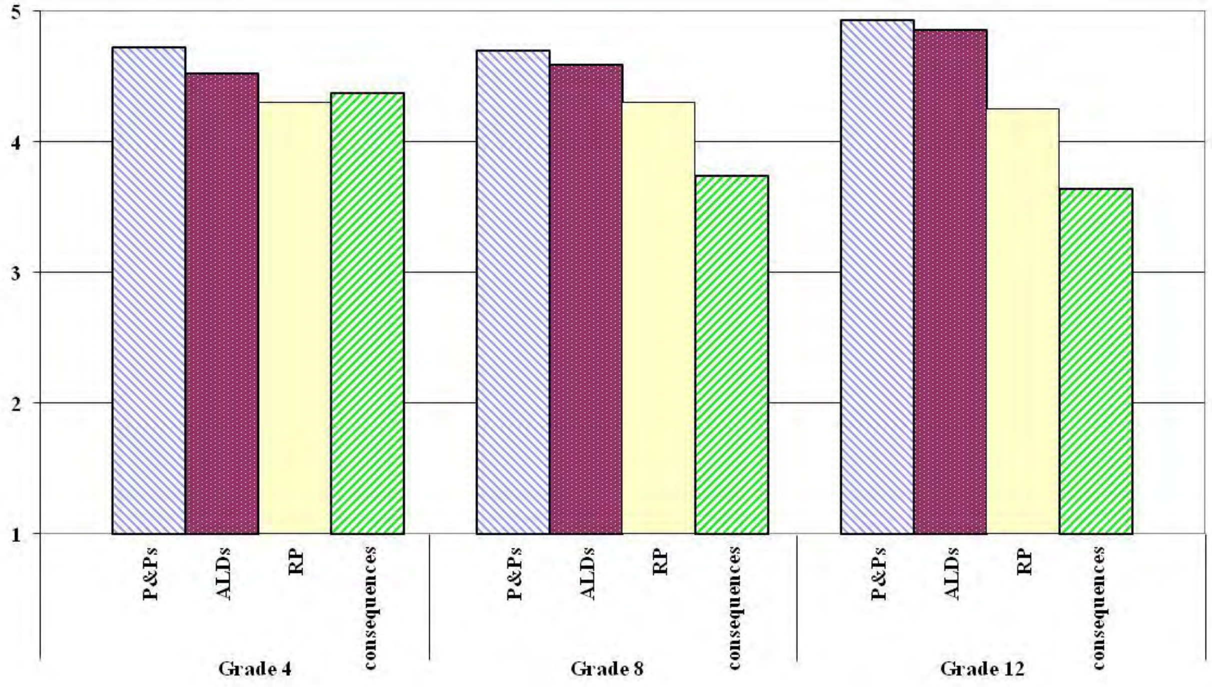


Figure 27: Average ratings of level of comfort with aspects, ALS round 3

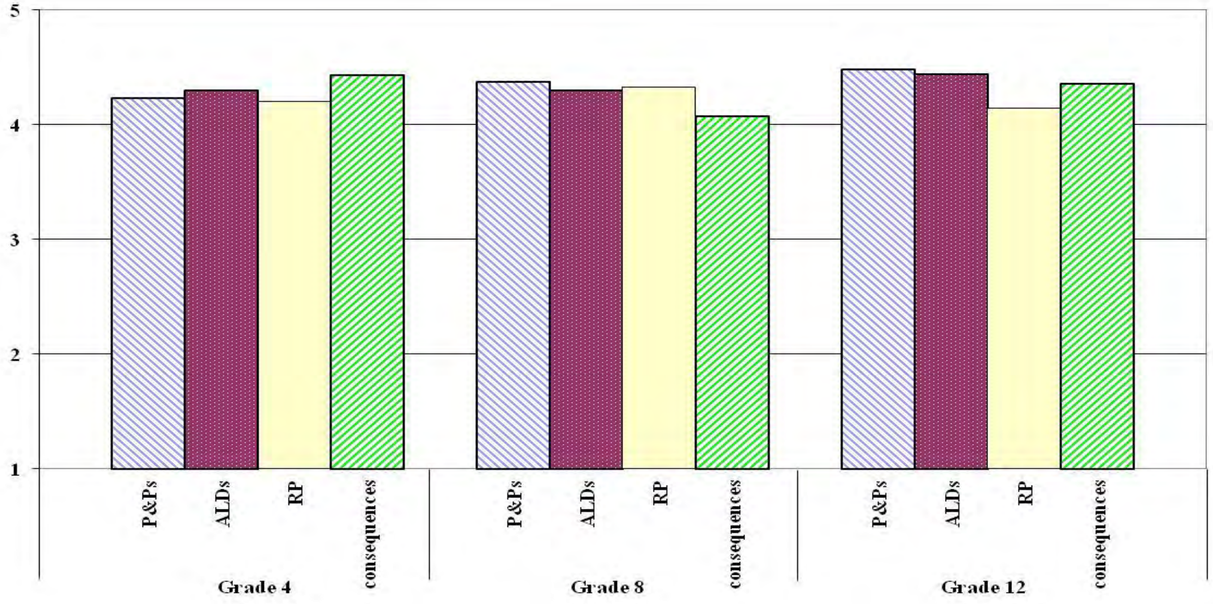


Figure 28: Average ratings of level of difficulty in using aspects, ALS round 3

The RP criteria tended to be ranked as more difficult than the other aspects of standard setting included in this research, and the science principles and practices tended to be ranked as the least difficult. The differences were not large and not consistent across grades. The average difficulty rankings (from 1 to 4, with 1 as the easiest and 4 as the most difficult) for each of the aspects and the percentage of panelists that ranked the topic as most difficult are shown in Tables 33 and 34 below.

Table 33: Average ranking of the difficulty of each aspect in ALS

Grade	Science Principles and Practices	Achievement Level Descriptions	Response Probability Criterion	Consequences Data
4	1.8	1.9	3.4	2.8
8	1.7	2.4	2.9	2.9
12	2.0	2.3	2.8	2.7

Table 34: Percentage ranking aspect as most difficult in ALS

Grade	Science Principles and Practices	Achievement Level Descriptions	Response Probability Criterion	Consequences Data
4	0	10	57	33
8	0	28	21	45
12	16	16	27	42

<u>Question 2</u>

To address this question, panelists were asked after round 2 in questionnaire 4 how a change in the RP value would affect their placement of the cut score. They were given the options of placing the bookmark earlier in the OIB, later in the OIB, or at the same place. The question was asked in two versions, one with an alternate RP value of 0.5 and one with an alternative RP value of 0.8, and these two versions of the question were alternated across panelists with half receiving each version.

To further address Question 2 at the end of the standard-setting process, the panelists were presented with two different statements describing possible approaches to thinking about what should happen with the bookmark when the RP value changes and asked which of the two was closer to their understanding of how things should work when the RP value is changed. The two statements are given below.

- **Statement A**: An ALD states what a student at a given achievement level should know and be able to do. The ALD for a given level provides the criteria for placing the bookmark for that level and is the same no matter what RP value is used. Therefore, if a different RP value is used, the same item should be bookmarked. Because of the new RP value, however, the scale value for the bookmarked item would change. Thus, a new cut score would be obtained.
- **Statement B**: A cut score for an achievement level represents the score value attained by a minimally-qualified student at that achievement level. The ALD for a given level

> provides the criteria for placing the bookmark for that level and is the same no matter what RP value is used. Therefore, if a different RP value is used, the score value attained by a minimally-qualified student (i.e., the cut score) should not change. Thus, a different item with the same, or very similar, scale score would become the bookmarked item.

These two statements represent different ways of thinking about how to choose a bookmark. The first statement represents the view that the ALDs define an achievement level, and that setting cut scores for an achievement level should only take into account these descriptions. This view requires panelists to pick the same item when the RP criterion changes. This implies that the cut scores would change when the RP value changes as the scale value associated with an item changes when the RP value changes. From this view, the RP criterion was used to rank order the items, and for nothing more.

The second statement represents the view that the task before the panelist is to rate the probability that a minimally-qualified student for an achievement level would get a point for each item. This view is in accord with the psychometric theory underlying the Bookmark method, which asserts that the cut score associated with an achievement level should not change when the RP criterion changes. In order to keep the same cut score, this view requires panelists to pick a different item when the RP criterion changes since the scale value associated with an item changes when the RP value changes.

For Question 2, the results indicate clearly that panelists struggle with notion of the RP criterion. The first evaluation question that dealt with this question asked how the panelists would change their cut score if the RP criterion were changed. If the RP criterion were changed to 0.5, the “correct” change would be to choose a more difficult item, that is, one farther back in the OIB. If the RP criterion were changed to 0.8, the panelist should choose an easier item, closer to the front of the book. This assumes, of course, that panelists are actually rating each item according to the probability that a minimally-qualified student for an achievement level would get that item correct. The results were similar for the two RP values. Table 35 shows the responses of ALS panelists by grade when responses were recorded to reflect the “correct” change.

Table 35: Percentage of ALS panelists changing cut score when RP criterion was changed

Grade	Correct Direction	Unchanged	Incorrect Direction	No Response
4	27	13	57	3
8	30	32	44	0
12	18	29	50	4
Total	25	22	51	2

For the question regarding changes in the cut score in response to changes in RP value (Statements A and B provided earlier), the panelists were essentially split. The results are shown in Table 36 for the ALS meeting. Logically, panelists who endorse Statement A should have answered the prior question by saying they would leave the cut score unchanged when the RP criterion changed, but this pattern did not hold generally.

Table 36: Percentage of ALS panelists endorsing Statements A and B

Grade	Statement A	Statement B	No Response
4	43	47	10
8	78	18	4
12	57	36	7
Total	59	34	7

Question 3

This topic was investigated after the exemplar item rating exercise. Since this was the last session of the ALS meeting, we thought it feasible to ask an open-ended question that could be followed by discussion during the debriefing. The question involved the distinction between mastery of the knowledge and skills in the ALD of a student at the proficient level, for example, and mastery of the knowledge and skills represented by the cut score set to represent the minimally-qualified student at the Proficient level. The question proposed a specific item as a potential exemplar that was in the middle of the Proficient achievement level. The probability of a correct response for the item was given for a student at the cut score. Since the item was above the cut score, this probability was below 0.67 and was close to 0.5. The panelists were then asked to explain how this could be considered a good example of performance at the Proficient level, when the probability was less than 0.67.

For Question 3, panelists were asked to respond to the following open-ended question on questionnaire 5:

> Item XX was chosen as a potential exemplar item for the proficient level. For this question, the chance that the minimally-qualified student at the Proficient cut score of YYY answers this item correctly is (some number less than 0.67). How would you explain to someone outside this panel why this is a good example of performance at the Proficient level, when the probability of getting a correct answer is less than 0.67 for the minimally-qualified student?

The responses for this question inform the question, do the panelists understand that the ALDs are for a range of student performances, and the cut score is for the lowest performance in that level? The responses were coded into one of three categories a correct response, a response that was centered on the primacy of the ALDs relative to the RP criterion, and any other type of response. Table 37 below shows the percentage of ALS panelists in each response category.

Table 37: Percentage of ALS panelists giving various explanations for why an item is a good example of performance

Grade	Correct Explanation	ALD Type Explanation	Other Explanation	No Response
4	63	20	13	3
8	41	26	33	0
12	43	29	25	4
Total	48	25	25	2

Note that the ALD type of explanation would be appropriate for panelists who are not considering the RP criterion as they set the cut score. While it might appear to be a response that does not really answer the question of interest, it would be consistent with panelists who endorsed Statement A. The explanations in the other category would typically mention that there were other reasons for students to miss the question (e.g., effort or time constraints). This does not really go to the heart of the question and might indicate a misunderstanding of the concept involved.

Outcomes of the Achievement Level Setting Process

There are three components of NAEP achievement levels: ALDs, cut scores, and exemplar items. The previous sections described the overall ALS process and the ALS meeting, which concern all three components. This section presents ACT's recommendations and information specific to each of the three components.

Achievement Level Descriptions

The ALDs represent the Governing Board's attempt to "stipulate what students should know and be able to do at each grade level and content area measured by NAEP" and to "make the NAEP data more understandable to the general user, parents, policymakers, and educators alike" (National Assessment Governing Board, 2008b). The ALDs were developed by the Governing Board before the ALS meeting (see Appendix C) and were translated into cut scores during the meeting.

On process evaluation questions in both the Pilot Study and in the ALS meeting, panelists reported being satisfied with the ALDs. Table 38 summarizes panelists' responses to questions concerning their satisfaction with the ALDs. Mean ratings of satisfaction with ALDs is consistently above 4 on a scale of 1–5, except for grade 12 in which there was some concern expressed about the description for life science at the Advanced level.

Table 38: Average ratings of responses to questions about ALDs

			ALS Average Rating		
Round	**Question #**	**Question**	**Grade 4**	**Grade 8**	**Grade 12**
1	2-17	The ALDs appear to be reasonably complete and comprehensive statements of what students should know and be able to do at each level of achievement.	4.00	4.15	3.58
1	2-18	My own level of satisfaction with the Basic ALD is:	4.27	4.07	4.12
1	2-19	My own level of satisfaction with the Proficient ALD is:	4.30	4.26	4.08
1	2-20	My own level of satisfaction with the Advanced ALD is:	4.33	4.33	3.78
Post	5-26	I believe that the achievement levels capture meaningful distinctions in science performance as described in the ALDs.	4.47	4.56	4.54

Panelists' average rating of their understanding of the ALDs is presented by level and round in Table 39. This question is asked for each level only at the conclusion of the first round, and is asked for all levels combined in succeeding rounds. At the conclusion of the first round, panelists were asked to indicate on a scale from 1–5 (1 = *totally inadequate*, 5 = *totally inadequate*): At the time I provided the round 1 bookmark placements, my understanding of the Basic/Proficient/Advanced ALD was ___. At the conclusion of the second and third rounds, the same question was asked for all three levels combined. Understanding of the ALDs could conceivably be viewed as an evaluation of the process, as opposed to the ALDs specifically. But panelists' understanding of the ALDs also reflects on how well the ALDs themselves can be understood by teachers, educators, and the general public. As shown in Table 39, panelists reported levels of understanding above 4.0 early in the process, and understanding increased noticeably over rounds as panelists continued to study and apply the ALDs to their tasks.

Table 39: Average ratings for understanding of ALDs

At the time I provided the/my round ___ bookmark placements/cut score recommendations my understanding of the _____ ALD was . . .

(5 = *totally adequate*; 3 = *somewhat adequate*; 1 = *totally inadequate*)

	ALS Average Rating				
Grade	**Basic Round 1**	**Proficient Round 1**	**Advanced Round 1**	**Round 2**	**Round 3**
4	4.03	4.07	4.03	4.66	4.80
8	4.11	4.26	4.19	4.59	4.70
12	4.41	4.37	4.33	4.63	4.89

As these ALDs were used to anchor the process for establishing cut scores and, as the responses of panelists to process evaluation questions concerning the ALDs are positive, ACT endorses the ALDs for use in representing the achievement levels set in this project.

Cut Scores

The cut scores from the ALS meeting are summarized by rater group, table group, and round for each grade in Table 40. (Cut score results for each panelist are contained in Appendix M.) The values in the rows labeled *Total* are the grade-group medians. The grade-group median is the cut score that was reported for each round. ACT recommends the round 3 medians, highlighted in yellow in Table 40, as the cut scores for the achievement levels. These numbers are on the ACT NAEP-like scale used in the ALS meeting.

Table 40: Science ALS cut scores* for each grade by rater group and table

Grade	Group		Basic			Proficient			Advanced		
			R1	R2	R3	R1	R2	R3	R1	R2	R3
4	Total		328	330	328	366	376	376	444	447	447
	Rater	A	328	330	328	373	384	380	444	444	444
		B	325	329	328	364	376	376	443	447	447
	Table	1	331	337	336	385	399	387	423	460	462
		2	320	327	327	359	363	363	444	440	435
		3	337	337	337	386	386	386	462	444	444
		4	296	315	326	351	362	367	424	435	445
		5	347	346	346	367	376	376	455	461	461
		6	325	328	328	351	378	378	443	447	447
8	Total		564	570	570	594	599	598	652	652	647
	Rater	A	563	566	566	599	598	598	657	658	653
		B	566	571	571	591	599	599	640	642	640
	Table	1	566	570	573	592	599	599	660	663	660
		2	564	570	570	598	601	598	657	661	647
		3	562	565	565	600	596	594	652	653	652
		4	563	571	571	604	604	604	633	647	638
		5	569	571	571	591	599	599	639	640	639
		6	565	565	568	593	591	597	668	647	646
12	Total		782	785	785	815	820	820	868	868	866
	Rater	A	775	784	784	815	822	822	867	867	865
		B	785	786	786	815	817	818	877	870	870
	Table	1	775	785	785	813	818	818	839	860	861
		2	755	780	780	809	823	823	868	871	871
		3	778	782	782	821	822	822	868	868	863
		4	786	787	787	824	832	828	886	877	877
		5	787	785	784	810	815	815	886	869	864
		6	783	784	784	816	816	817	859	867	867

*The ACT NAEP-like scales have means (SDs) of 364 (33), 579 (33), and 793 (33) for grades 4, 8, and 12, respectively.

ACT conducted extensive statistical analysis on the cut scores in order to assess characteristics related to the reliability of the medians for each grade and the overall quality of the ALS process. Key analyses and conclusions are summarized in the next two sections.

Distribution of Cut Scores by Round

The variability of cut scores within rounds, levels, and grades was assessed. The median is typically used in bookmark-based methods because the median is less sensitive to outliers than the mean. It is relatively easy for a bookmark or Mapmark panelist to provide an extreme cut score recommendation either out of inexperience or in an attempt to influence the mean. As panelists review results and feedback together, outliers and variability tend to decrease as panelists gain a shared sense of borderline performance and as they become aware of the group cut score. In all but one case (grade 12, Advanced level), the variability of cut scores across panelists in the 2009 Science ALS decreased by round. The variability of the cut scores was approximately the same for rounds 2 and 3 for grade 12 at the Advanced level. For each grade, plots of the Mean Absolute Deviation (MAD) of cut scores of individual panelists from the group cut score by round in the ALS are shown in Figures 29–31. For grades 8 and 12, the ALS MAD was largest for the Advanced level in round 1 and then decreased in subsequent rounds. For grade 4, the ALS MAD for the Advanced and Basic levels were similar in round 1 and then decreased in subsequent rounds. Variation in panelists' cut score recommendations decreased over rounds with the greatest amount of convergence between rounds 1 and 2. In addition, the lack of large increases in the MAD from round 2 to 3 indicates that there were no extreme reactions among panelists to student performance data in the ALS. These findings are consistent with results ACT has obtained in previous standard-setting work for the Governing Board.

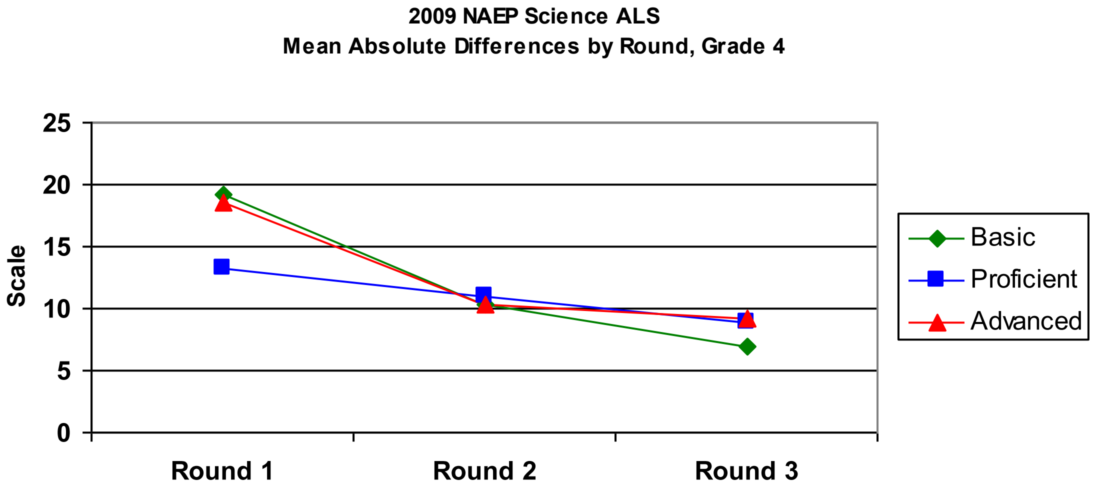


Figure 29: Mean Absolute Deviation (MAD) of cut scores from median by round for grade 4

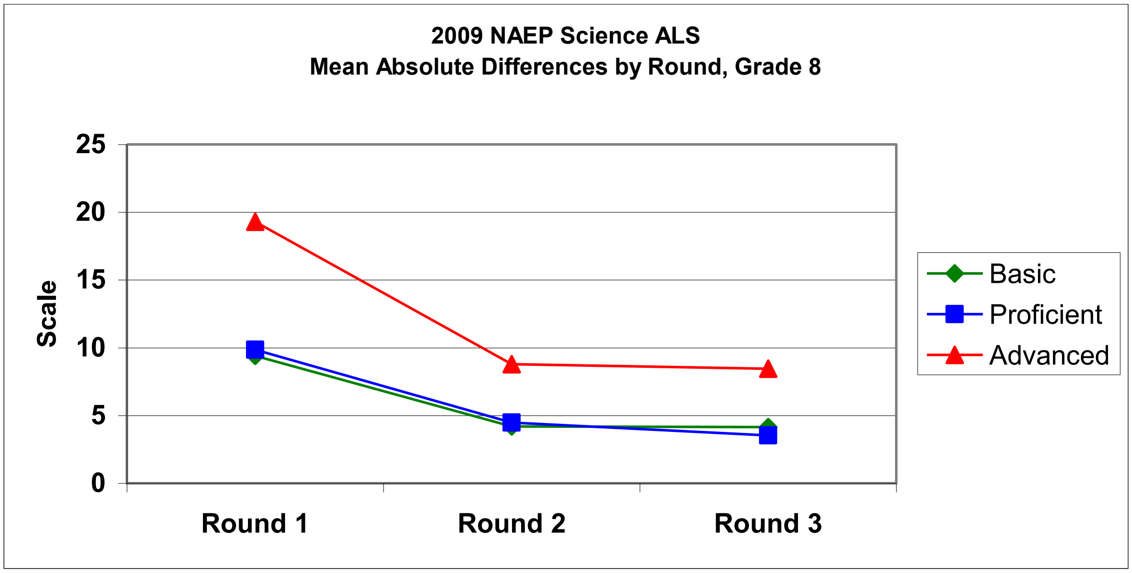


Figure 30: Mean Absolute Deviation (MAD) of cut scores from median by round for grade 8

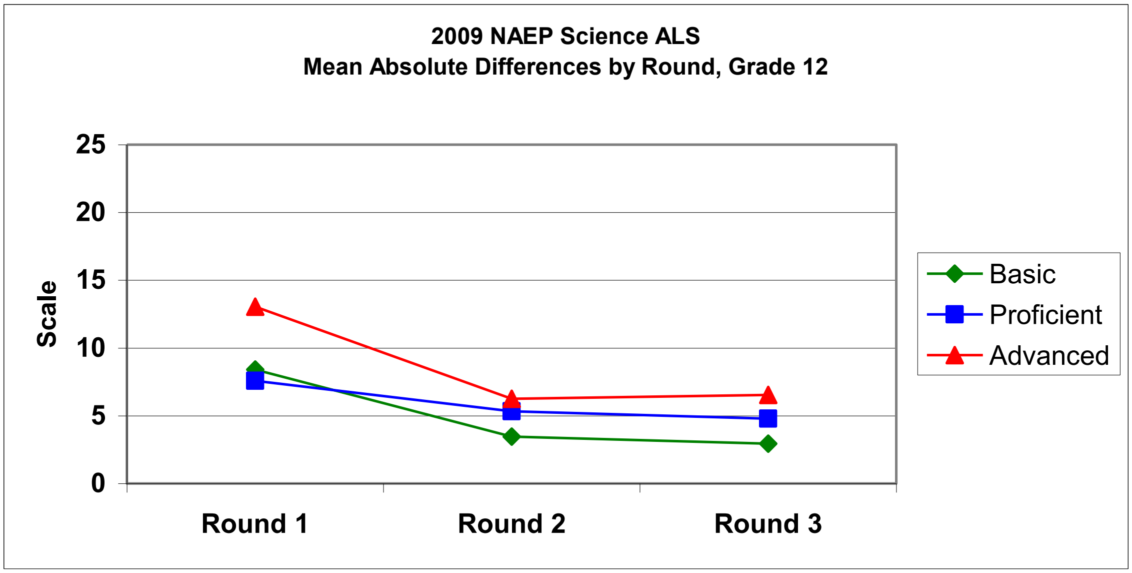


Figure 31: Mean Absolute Deviation (MAD) of cut scores from median by round for grade 12

A study of the change in cut scores by level and round provides additional information about how panelists were responding to the feedback provided. Table 41 presents the number and percent of panelists whose cut scores increased from the previous round, decreased, or had no change. The patterns in this table are similar to the patterns seen in previous standard settings for the Governing Board. The largest frequency of change is from round 1 to 2, indicating the incorporation of information gleaned from the booklets and the cut scores of other panelists into their judgments. At each level, except Advanced for grade 8, the majority of panelists did not change their cut scores after round 2.

Table 41: Number and percent of panelists who changed their cut scores between rounds in the ALS

Changes Between Rounds	Basic			Proficient			Advanced		
	Increase N (%)	No Change N (%)	Decrease N (%)	Increase N (%)	No Change N (%)	Decrease N (%)	Increase N (%)	No Change N (%)	Decrease N (%)
Grade 4									
1 to 2	17 (57)	6 (20)	7 (23)	24 (80)	1 (3)	5 (17)	16 (53)	3 (10)	11 (37)
2 to 3	7 (23)	18 (60)	5 (17)	6 (20)	15 (50)	9 (30)	4 (13)	19 (63)	7 (23)
Grade 8									
1 to 2	16 (59)	5 (19)	6 (22)	14 (52)	4 (15)	9 (33)	14 (52)	3 (11)	10 (37)
2 to 3	3 (11)	20(74)	4 (15)	1 (4)	20 (74)	6 (22)	1 (4)	11 (41)	15 (56)
Grade 12									
1 to 2	16 (57)	9 (32)	3 (11)	16 (57)	6 (21)	6 (21)	13 (46)	9 (32)	6 (21)
2 to 3	1 (4)	26 (93)	1 (4)	2 (7)	24 (86)	2 (7)	1 (4)	17 (61)	10 (36)

As shown in Table 42, differences between the mean and median cut scores were generally small. The largest difference was five points at the grade 4 Basic level in round 3. In this case and in the cases of the grade 4 Advanced level in rounds 1 and 2, the highest cut scores recommended by the panelists tended to be higher than one would expect in a symmetrical distribution of cut scores. In the case of the grade 12 Basic level in round 1 the opposite is true; the lowest cut scores recommended by the panelists tended to be lower than one would expect in a symmetrical distribution of cut scores.

The variation in the differences of the mean and median (some larger positive numbers and one smaller negative number) may have been due to the sparseness of the items at the low end and at the high end of the scale for each grade. At the individual level, a change of one or two items from Basic to Below Basic corresponded to a relatively large increase in the scale score selected as the Basic cut score, and a change of one or two items from Advanced to Proficient corresponded to a relatively large increase in the scale score selected as the Advanced cut score. Thus, individuals could reasonably vary the values of their Basic and Advanced cut scores, contributing to variations in the mean and median differences at these levels.

Table 42: Mean and median cut scores and difference by round for ALS

	Achievement	Mean			Median			Mean - Median		
Grade	Level	R1	R2	R3	R1	R2	R3	R1	R2	R3
4	Basic	325	333	333	328	330	328	-3	3	5
	Proficient	365	377	375	366	376	376	-1	1	-1
	Advanced	448	451	450	444	447	447	4	4	3
8	Basic	564	569	569	564	570	570	0	-1	-1
	Proficient	595	598	598	594	599	598	1	-1	0
	Advanced	655	652	648	652	652	647	3	0	1
12	Basic	778	786	785	782	785	785	-4	1	0
	Proficient	817	820	820	815	820	820	2	0	0
	Advanced	868	869	867	868	868	866	0	1	1

Reliability of Cut Scores

The reliability of cut scores emerging from a standard-setting process is typically thought of in regard to how consistent the cut scores are across tables, rater groups, and panelist type, and how close the final cut scores from the process would be if the process were performed on two occasions with few differences.

After a thorough review of the effects of design factors (tables and groups) and panelist characteristics on cut scores, ACT's Technical Advisory Committee on Standard Setting did not identify any effects that called the results of the ALS meeting into question or raised serious questions about the process.

As there is no satisfactory method of estimating the significance of the differences between groups on their median cut scores and as the mean and median cut scores were highly similar, ACT performed analyses of the effects on means. Very few statistically significant effects emerged from the analyses for panelists' gender, race/ethnicity, geographic region or type. Several statistically significant effects appeared for the analyses of tables and rater groups. The differences that were statistically significant will be mentioned along with a brief description of differences in medians.

Rater group (item pool) effects were statistically significant for grade 4 Proficient level in round 1, for grade 8 Advanced level in the final round, for grade 12 Basic level in round 1 and 3, and for grade 12 Advanced level in round 1. Figures 32–34 show the medians for each rater group by round and achievement level.

The graph in Figure 33 of the grade 8 rater-group (item pool) effects based on the group median cut scores at the Advanced level would seem to illustrate that table differences were similar, but somewhat larger, at round 1 than round 3 (16.5 and 13, respectively, from Table 44). However, variance in the first round was greater within than between rater groups, whereas by the final round, the variance had decreased substantially so as to render smaller mean differences significant (see Appendix M for cut scores by panelist within groups and tables). Note that there had been a lot of table discussion of the results by the end of round 3, and the assumption of independence for tests of differences may be questionable at this stage.

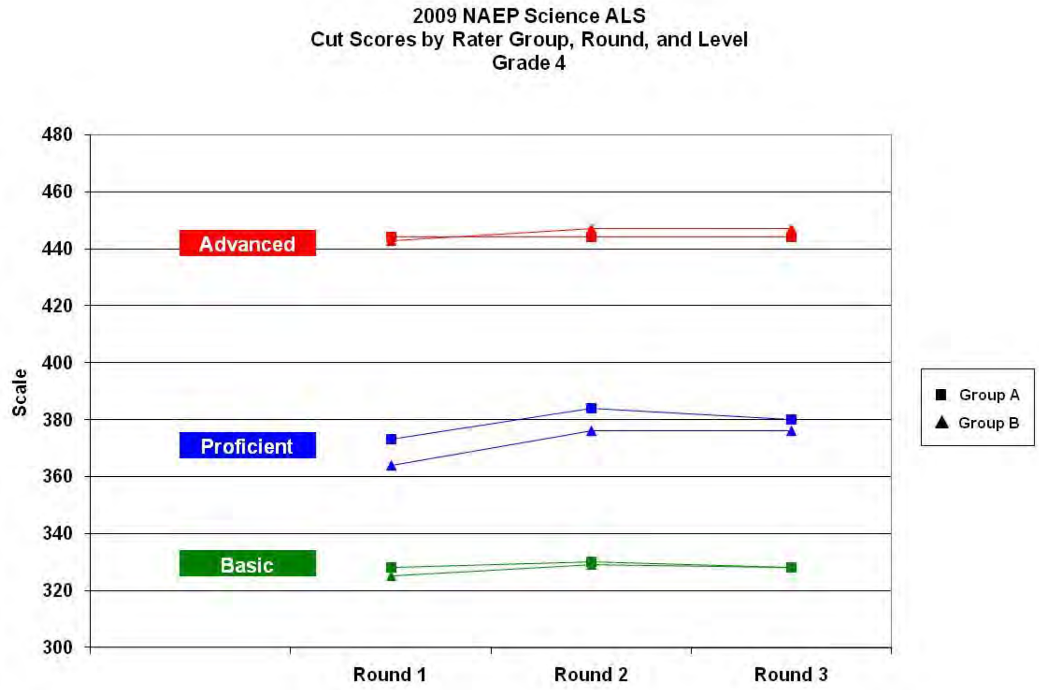


Figure 32: Median cut scores by item pool/rater group, round, and level for grade 4

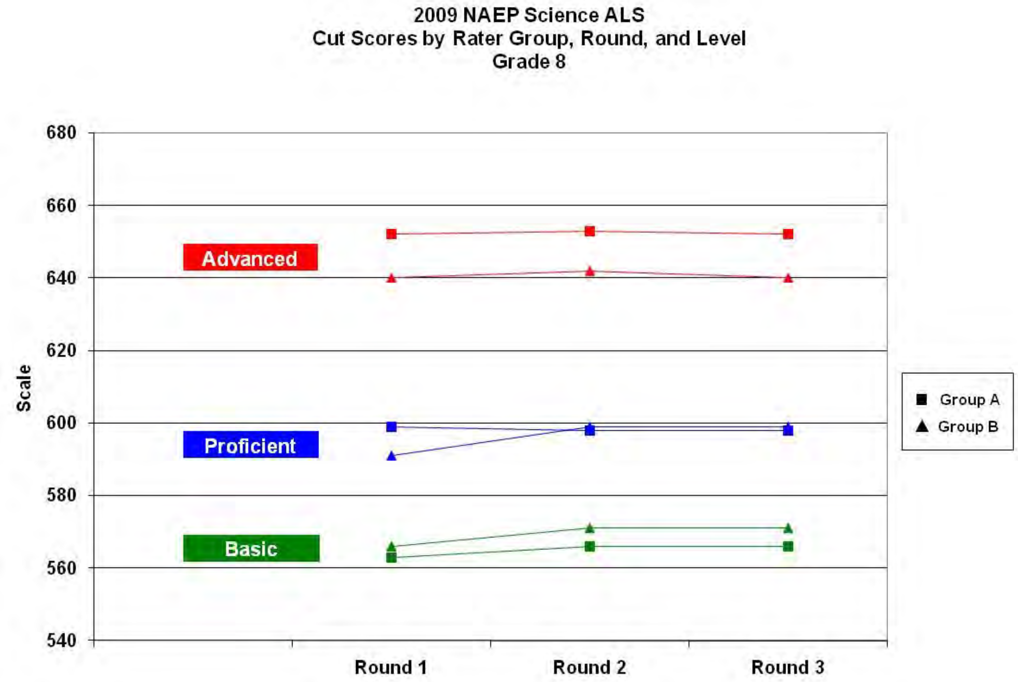


Figure 33: Median cut scores by item pool/rater group, round, and level for grade 8

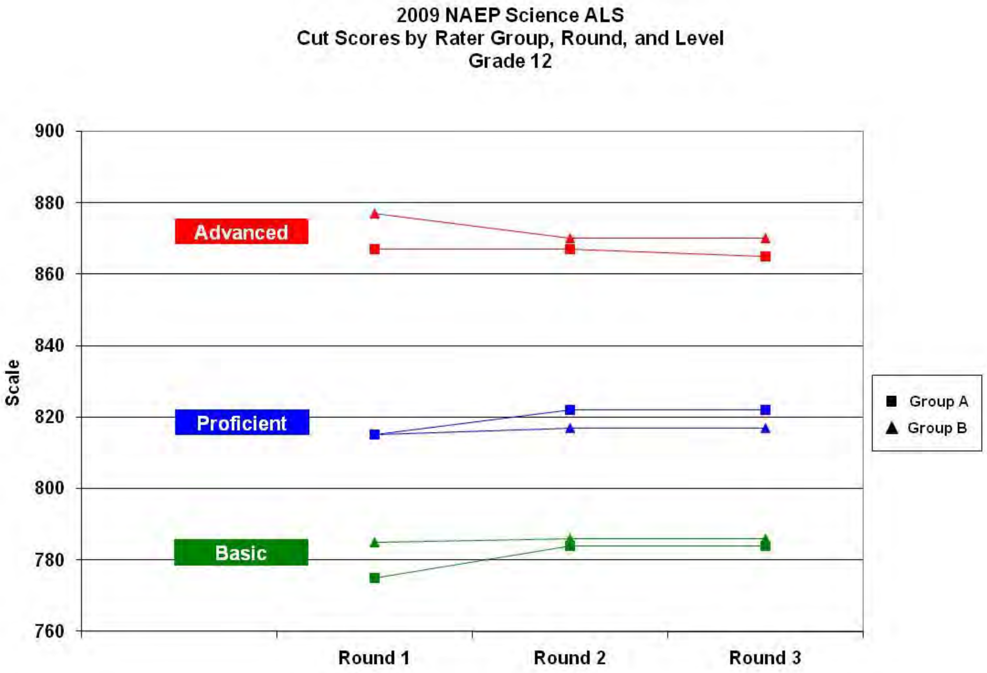


Figure 34: Median cut scores by item pool/rater group, round, and level for grade 12

Figures 35–37 show table medians by round and achievement level. Grade 4 table group effects were statistically significant at both the Basic and Proficient levels for round 1 and at all three achievement levels for round 3. Grade 8 table group effects were statistically significant at the Advanced level for the final round. Grade12 table group effects were statistically significant at the Basic level for rounds 1 and 3, at the Proficient level for round 3 and for the Advanced level for rounds 1 and 3 (see Appendix M for cut scores by panelist within groups and tables). Tables 43–45 show that the largest round 1 within-group difference was 51 points between table median Basic cut scores at grade 4 and the largest final round within-group difference was 27 points between table median Advanced cut scores at grade 4. Note that some table groups are quite small and the average values could be affected by a single outlier. This could lead to significant differences in round 1. At round 3, again the differences within a table are usually very small, and as a result, even small differences would be statistically significant.

Finally, differences in cut scores between different genders, races/ethnicities, geographic regions, and panelist types (teacher, nonteacher, general public) were not statistically significant, except between regions at the Proficient and Advanced levels for the final round at grade 8. Given the number of comparisons made, and the small sample size for some of the groups, it is not unusual to find a few significant results. Table 44 shows that the median differences that correspond to the statistically significant mean differences were 7 and 13, respectively. The median differences between regions at the Proficient and Advanced levels for the first round at grade 8 were 13.5 and 8, respectively.

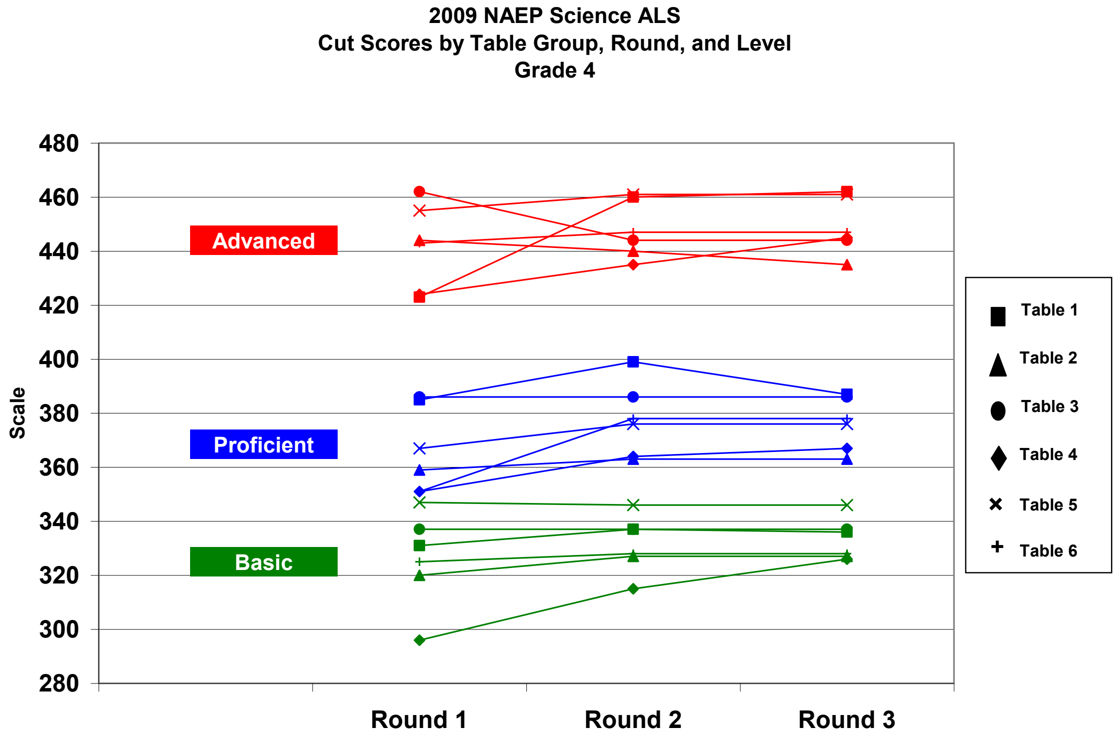


Figure 35: Median cut scores by table, round, and achievement level for grade 4

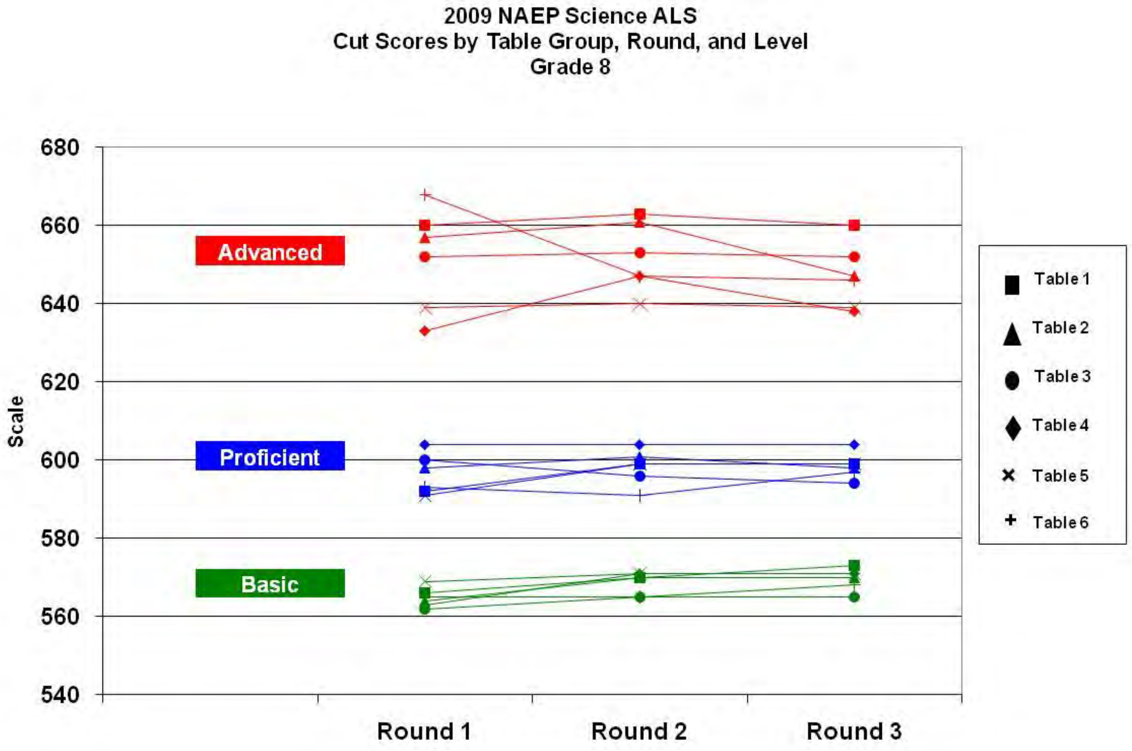


Figure 36: Median cut scores by table, round, and achievement level for grade 8

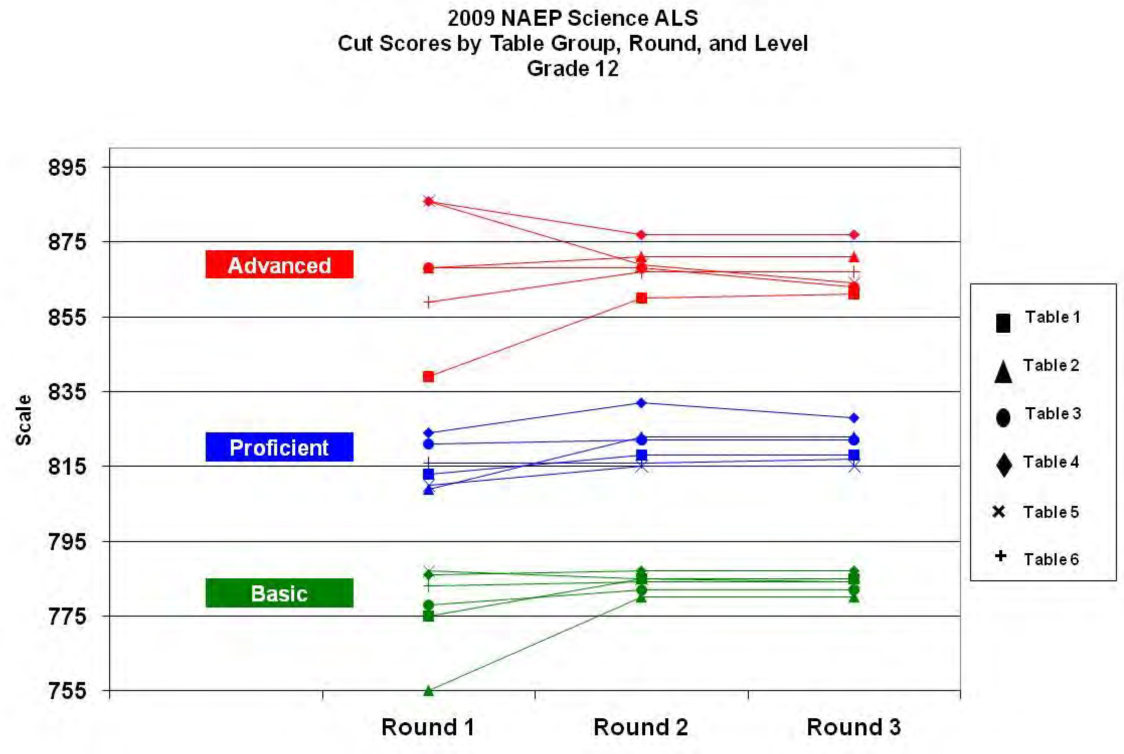


Figure 37: Median cut scores by table, round, and achievement level for grade 12

Table 43: Grade 4 medians and mean absolute difference (MAD) of cut scores by factor level

		Round 1						Round 3					
		Basic		Proficient		Advanced		Basic		Proficient		Advanced	
Factor	N	Median	MAD	Median	MAD	Median	MAD	Median	MAD	Median	MAD	Median	MAD
Group													
A	15	328	14.2	373	14.3	444	22.5	328	6.8	380	12.8	444	11.1
B	15	325	23.9	364	11.5	443	14.1	328	7.1	376	3.8	447	6.9
Difference		3		9		1		0		4		3	
Table													
A-1	5	331	17.6	385	7.0	423	36.0	336	6.0	387	3.6	462	9.8
A-2	5	320	8.4	359	1.4	444	15.2	327	2.0	363	7.0	435	1.8
A-3	5	337	12.2	386	13.0	462	8.6	337	8.0	386	5.8	444	6.2
Max. Diff.		17		27		39		10		24		27	
B-4	5	296	29.6	351	16.2	424	14.8	326	1.6	367	1.8	445	3.6
B-5	5	347	2.4	367	0.0	455	11.0	346	0.0	376	0.4	461	0.2
B-6	5	325	9.4	351	10.2	443	5.6	328	1.0	378	0.6	447	2.4
Max. Diff.		51		16		31		20		11		16	
Type													
Teacher	16	326	20.6	364	12.4	443.5	17.6	328.5	5.6	376	8.3	447	10.1
Nonteacher	5	353	10.2	367	17.6	443	12.8	346	8.6	378	9.2	444	5.4
Gen. Public	9	318	12.7	367	11.6	445	23.1	328	6.6	376	9.3	447	9.0
Max. Diff.		35		3		2		18		2		3	
Region													
Midwest	7	328	27.4	362	21.6	443	14.6	328	9.4	378	10.7	444	5.6
Northeast	5	314	11.2	364	15.0	444	16.6	325	5.2	376	8.2	447	12.4
South	11	325	17.9	367	8.5	459	22.2	328	4.5	376	7.3	447	9.0
West	7	337	9.9	367	9.7	435	11.0	337	7.6	378	9.3	447	10.1
Max. Diff.		23		5		24		12		2		3	
Race/Ethnicity													
Nonminority	26	326	15.3	365.5	11.8	443.5	18.4	328	6.9	376	8.4	447	9.6
Minority	4	356	31.3	378.5	22.3	461	11.5	332.5	7.5	374.5	11.5	445.5	7.0
Difference		30		13		17.5		4.5		1.5		1.5	
Gender													
Male	12	325	13.8	367	11.1	450	23.6	328	7.6	376.5	9.3	446	10.8
Female	18	329.5	22.6	363	13.9	443.5	15.0	329	6.4	376	8.5	447	8.3
Difference		4.5		4		6.5		1		0.5		1	

Table 44: Grade 8 medians and mean absolute difference (MAD) of cut scores by factor level

		Round 1						Round 3					
		Basic		Proficient		Advanced		Basic		Proficient		Advanced	
Factor	**N**	**Median**	**MAD**	**Median**	**MAD**	**Median**	**MAD**	**Median**	**MAD**	**Median**	**MAD**	**Median**	**MAD**
Group													
A	14	563	8.8	598.5	10.4	656.5	19.6	566	4.5	598	2.9	653	5.7
B	13	566	9.9	591	8.5	640	13.9	571	3.1	599	4.1	640	5.5
Difference		3		7.5		16.5		5		1		13	
Table													
A-1	4	565.5	9.5	591.5	10.3	660	8.5	573	5.5	599	3.3	660	5.3
A-2	5	564	9.6	598	10.2	657	18.4	570	4.0	598	2.0	647	7.6
A-3	5	562	7.0	600	10.2	652	28.6	565	1.6	594	1.6	652	1.2
Max. Diff.		3.5		8.5		8		8		5		13	
B-4	5	563	18.0	604	7.2	633	4.0	571	4.8	604	4.8	638	3.0
B-5	4	569	2.5	591	5.3	639	6.5	571	0.5	599	1.3	639	2.5
B-6	4	565	5.0	592.5	10.0	667.5	26.0	568	3.0	597	3.3	646	9.5
Max. Diff.		6		13		34.5		3		7		8	
Type													
Teacher	13	566	10.6	594	9.8	647	21.9	570	4.8	599	4.0	643	6.7
Nonteacher	6	564	9.2	594.5	8.5	655	19.3	567	2.5	598	2.0	657	9.0
Gen. Public	8	562	6.6	595	11.0	647.5	14.1	571	3.4	598	3.8	647	8.1
Max. Diff.		4		1		8		4		1		14	
Region													
Midwest	6	561	6.83	591	7.0	654.5	21.7	565	3.5	592	5.0	643	10.3
Northeast	7	561	13.0	590	11.7	647	17.3	567	3.4	598	1.9	656	5.6
South	8	564.5	8.6	594	9.8	647	17.4	571	4.8	599	3.4	643	3.8
West	6	567.5	6.0	603.5	7.7	646.5	20.7	571	2.3	599	1.8	646	8.7
Max. Diff.		6.5		13.5		8		6		7		13	
Race/Ethnicity													
Nonminority	23	566	8.7	594	11.0	652	19.2	570	4.5	598	3.8	647	9.0
Minority	4	562.5	13.0	596	3.5	649.5	19.8	569	2.3	598	1.8	650	5.0
Difference		3.5		2		2.5		1		0		3	
Gender													
Male	13	564	10.6	591	9.5	643	16.2	570	2.8	598	2.5	647	8.4
Female	14	565	8.3	598.5	8.5	654.5	20.6	570	5.4	598	4.5	646	8.5
Difference		1		7.5		11.5		0		0		1	

Table 45: Grade 12 medians and mean absolute difference (MAD) of cut scores by factor level

		Round 1						Round 3					
		Basic		Proficient		Advanced		Basic		Proficient		Advanced	
Factor	N	Median	MAD	Median	MAD	Median	MAD	Median	MAD	Median	MAD	Median	MAD
Group													
A	14	775	8.4	822	9.0	867	8.4	784	2.5	815	3.4	865	4.1
B	14	785	4.4	818	6.1	876.5	16.1	786	3.2	815	5.6	870	8.2
Difference		10		4		9.5		2		0		5	
Table													
A-1	4	775	4.0	813	3.0	838.5	10.3	785	0.3	818	3.3	861	4.8
A-2	5	755	11.4	809	14.0	868	2.8	780	3.4	823	3.8	871	3.8
A-3	5	778	4.4	821	5.2	868	2.0	782	1.0	822	1.4	863	1.2
Max. Diff.		23		12		29.5		5		5		10	
B-4	5	786	1.0	824	6.0	886	10.4	787	4.4	828	6.2	877	5.0
B-5	5	787	9.0	810	3.0	886	16.2	784	2.0	815	1.0	864	8.0
B-6	4	783	1.5	816	6.3	859	9.8	784	1.0	817	2.8	867	5.3
Max. Diff.		4		14		27		3		13		13	
Type													
Teacher	17	782	10.6	814	5.1	868	11.9	785	3.5	818	5.1	866	6.8
Nonteacher	5	782	4.2	820	8.8	866	20.6	785	1.6	819	3.4	868	6.8
Gen. Public	6	782	5.5	821	11.2	869.5	9.5	784	2.2	822.5	3.8	868	5.2
Max. Diff.		0		7		3.5		1		4.5		2	
Region													
Midwest	7	776	8.3	819	6.7	868	15.0	782	2.1	821	4.7	866	5.9
Northeast	5	784	4.0	815	10.0	868	8.4	784	4.0	823	6.4	871	5.6
South	9	784	8.0	814	5.9	866	12.3	785	3.0	819	3.7	866	6.2
West	7	782	9.6	810	6.6	868	15.0	785	2.0	819	4.3	866	7.6
Max. Diff.		8		9		2		3		4		5	
Race/Ethnicity													
Nonminority	21	778	8.8	815	8.3	868	13.7	784	2.5	821	5.0	866	6.6
Minority	7	784	5.7	815	5.3	868	11.1	786	3.7	818	3.6	869	6.0
Difference		6		0		0		2		3		3	
Gender													
Male	17	778	6.5	819	7.5	868	14.2	784	2.5	821	4.9	866	7.2
Female	11	784	10.5	812	7.1	868	11.2	785	3.5	819	4.5	868	5.4
Difference		6		7		0		1		2		2	

Cut scores by round for the Pilot Study and ALS are presented in Table 46. On the NAEP-like scales, the final Basic cut scores from the two meetings differed by 1, 9, and 9 points, the final Proficient cut scores differed by 10, 1, and 8 points, and the final Advanced cut scores differed by 11, 14, and 4 points for grades 4, 8, and 12, respectively.

Table 46: Cut scores by grade by round and achievement level for Pilot Study and ALS*

	Basic			Proficient			Advanced		
Study	**Round 1**	**Round 2**	**Round 3**	**Round 1**	**Round 2**	**Round 3**	**Round 1**	**Round 2**	**Round 3**
Grade 4									
ALS	328	330	328	366	376	376	444	447	447
Pilot	316	325	327	362	366	366	436	436	436
Difference	12	5	1	4	10	10	8	11	11
Grade 8									
ALS	564	570	570	594	599	598	652	652	647
Pilot	561	561	561	598	599	599	659	661	661
Difference	3	9	9	4	0	1	7	9	14
Grade 12									
ALS	782	785	785	815	820	820	868	868	866
Pilot	772	776	776	811	812	812	875	871	870
Difference	10	9	9	4	8	8	7	3	4

*The ACT NAEP-like scales have means (SDs) of 365 (33), 579 (33), and 793 (33) for grades 4, 8, and 12, respectively.

The standard error of the cut score is an estimate of the uncertainty in the reported cut score (the median cut score across panelists) due to various sources of error. The standard error of the difference of two cut scores combines the estimates of the standard error of each individual cut score. Unfortunately, ACT can recommend no single, straightforward method for estimating the standard error of the final cut score in the typical standard-setting process in which panelists recommend cut scores over rounds, based in part on feedback they receive about the cut score from the previous round. Panelist cut scores after round 1 are influenced by the group cut scores and cut score distributions. Panelists are generally more comfortable being close to the middle so there is a regression to the round 1 group cut score for each level. Estimates of the standard error of the final cut score do not account for a fundamental regression to the median of previous rounds, motivated by panelists' desire for conformity, as well as for the effects of criterion-referenced feedback. For this reason, estimates of the standard error at the final round tend to be smaller and are more likely to underestimate differences between replications of a method using the same item pools but different groups of panelists. In addition, cut scores established in rounds 2 and 3 are based on the baseline established in the first round, and do not tend to vary substantially from the previous round. For this reason, an understanding of the differences between cut scores is most informed by an analysis of results from round 1.

Table 47 presents the standard error estimates for the group cut scores (medians) for round 1 and the final round for each achievement level at each grade in the Pilot Study and ALS, with the standard errors calculated using two distinct nonparametric methods (Maritz & Jarrett, 1978; bootstrap, see Efron & Gong, 1983). As expected, the standard errors generally decreased from round 1 to the final round. The standard errors of the difference between the ALS and Pilot Study cut scores are shown in Table 48 and are compared to the absolute values of the actual differences. The actual differences between the round 1 cut scores were close to one standard error of the difference for the Proficient and Advanced levels at all three grades and for the Basic level for grade 8. The differences between round 1 cut scores were close to two standard errors for the Basic level for grades 4 and 12. As estimates of the standard error at the final round are underestimates, the relevant round for interpretation of differences is the first round.

Table 47: Estimates of standard error of the cut scores across achievement level and round for the Pilot Study and ALS, using two distinct nonparametric methods

			Basic		Proficient		Advanced	
Study	**Grade**	**Statistical Method**	**Round 1**	**Final**	**Round 1**	**Final**	**Round 1**	**Final**
ALS	4	Maritz-Jarrett SE	5.2	2.0	2.2	2.0	3.8	3.8
		Bootstrap SE	4.9	2.0	2.1	1.6	4.0	3.7
	8	Maritz-Jarrett SE	2.2	1.5	3.3	0.8	6.0	3.6
		Bootstrap SE	2.2	1.4	3.1	0.8	5.6	3.5
	12	Maritz-Jarrett SE	2.5	0.7	2.2	1.5	1.9	2.0
		Bootstrap SE	2.1	0.6	2.2	1.4	1.6	2.0
Pilot Study	4	Maritz-Jarrett SE	3.8	3.5	1.9	3.2	11.9	3.8
		Bootstrap SE	3.6	3.4	1.7	3.1	11.3	3.4
	8	Maritz-Jarrett SE	4.7	2.4	5.0	1.7	7.4	2.0
		Bootstrap SE	4.2	1.7	4.8	1.0	6.4	1.7
	12	Maritz-Jarrett SE	4.0	2.6	5.8	3.3	5.9	2.2
		Bootstrap SE	3.0	1.2	4.6	1.7	4.8	1.4

Table 48: Estimates of standard error of the difference in the Pilot Study and ALS group cut scores by achievement level and round compared to absolute value of actual difference

		Basic		Proficient		Advanced	
Grade	**Standard Error of the Difference**	**Round 1**	**Final**	**Round 1**	**Final**	**Round 1**	**Final**
4	Maritz-Jarrett	6.4	4.0	3.0	3.8	12.5	5.4
	Bootstrap	6.1	3.9	2.7	3.5	12.0	5.0
	Observed |D|	12	1	4	10	8	11
8	Maritz-Jarrett	5.2	2.8	5.9	1.9	9.6	4.2
	Bootstrap	4.7	2.2	5.8	1.3	8.5	3.9
	Observed |D|	3	9	4	1	7	14
12	Maritz-Jarrett	4.7	2.7	6.2	3.6	6.2	3.0
	Bootstrap	3.6	1.4	5.1	2.2	5.1	2.4
	Observed |D|	10	9	4	8	7	4

Differences in cut scores may be due to factors expected to affect cut scores, which vary across meetings using the same method, but which are not represented in the standard error estimates. Such factors include physical accommodations, presence of observers, interactions among panelists over rounds, random variation, and panelist understanding of the purpose of the meeting. ACT and our Technical Advisory Committee on Standard Setting carefully reviewed procedural validity and internal consistency data from the ALS to determine if differences may have been due to procedural or internal validity factors. In addition, ACT reviewed panelists’ qualifications. Results indicated no differences in panelist qualifications between the Pilot Study and ALS, that the ALS procedural results were stronger than or comparable to that of the Pilot Study, and that internal consistency emerged as expected.

The conclusion was that there is no reason to doubt the results of the ALS, and that the differences between the results from the two meetings may be due to the panelists’ understanding of differences in the purpose of the meetings. Panelists in the Pilot Study clearly understood that the cut scores they established would not have national implications but, instead, would inform development and refinement of the method. In general, panelists in ALS meetings may be more likely to set lower cut scores because they know that the scores will be used for reporting the national results of student performance. Panelists in the Pilot Study knew the results would not be reported. This difference, along with the changes to the briefing book, the presentations, and the number of panelists, means that the ALS is not an exact replication of the Pilot Study.

Exemplar Item Ratings

Exemplar item ratings were gathered in the ALS meeting to provide the Governing Board with information concerning the suitability of assessment items for illustrating what students know and can do at each level of achievement.

Potential exemplar items were drawn from two blocks of the assessment selected for possible release to the public for each grade level. Items in these two blocks were common to both item-rating pools at each grade level and were shaded yellow on the Primary Item Map. Items/score points were mapped to the first, or easiest, achievement level at which the probability was 0.67 or higher that a student at the top of the level could correctly answer the item or attain the score point. For example, at the Proficient level, all items to be released that mapped to a value between the Proficient and Advanced cut scores were selected as potential exemplars for the Proficient level (see Figure 38). Recall that each score point of a constructed-response item was mapped independently of other score points by the probability of scoring at or above the score point. The number of score points per achievement level overall and by item type is shown in Table 49.

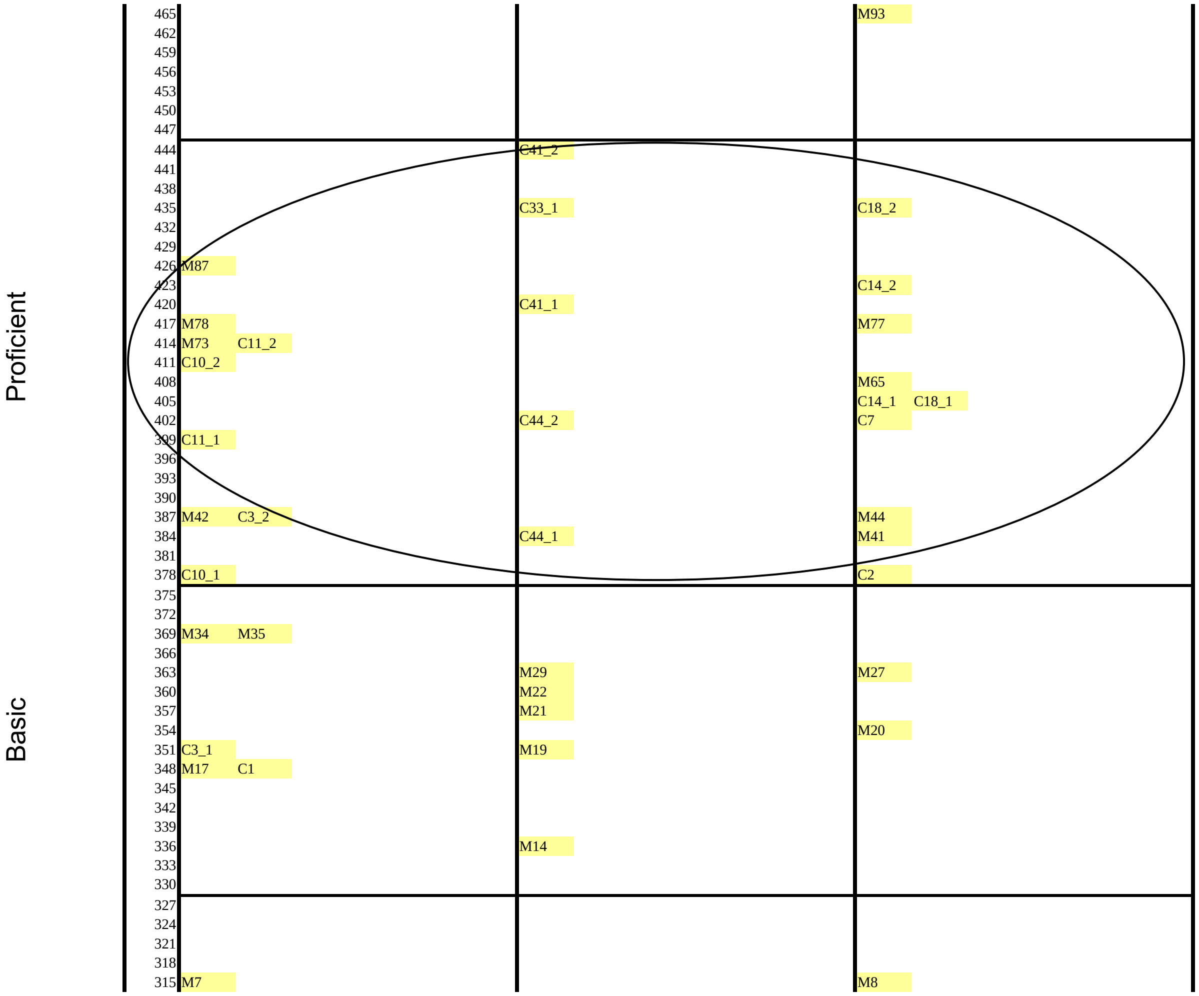


Figure 38: Potential exemplar items selected to represent Proficient level

Table 49: Number of multiple-choice items and constructed-response score points identified as potential exemplars

Grade	Achievement Level	Multiple Choice	Constructed Response	Total
4	Basic	10	2	12
	Proficient	8	16	24
	Advanced	1	4	5
8	Basic	5	6	11
	Proficient	10	13	23
	Advanced	1	11	12
12	Basic	7	3	10
	Proficient	11	11	22
	Advanced	1	10	11

For each item, panelists were asked to indicate if they felt the item was *Very Good*, *OK*, or *Do Not Use* to illustrate performance at the level with which it was associated. Detailed results of the exemplar item ratings are shown in Appendix D. ACT and our TACSS recommended that the Governing Board use items rated by 50% or more of the panelists as *Very Good* and by fewer than 30% of the panelists as *Do Not Use* as exemplars in the reporting of NAEP results. The shaded cells in Appendix D identify items that meet that criterion. The number of potential exemplar items/score points, per achievement level, meeting this criterion overall and by item type is provided in Table 50.

Table 50: Number of multiple-choice items and constructed-response score points recommended for use as exemplars

Grade	Achievement Level	Multiple Choice	Constructed Response	Total
4	Basic	2	2	4
	Proficient	5	4	9
	Advanced	0	4	4
8	Basic	3	4	7
	Proficient	1	5	6
	Advanced	0	5	5
12	Basic	3	0	3
	Proficient	4	5	9
	Advanced	0	7	7

Each achievement level, except for grade 12 Basic, was associated with at least two score points on constructed-response items that met ACT's suggested ratings criteria. There was only 1 MC item eligible for selection as an exemplar for each Advanced level, and none met the rating criteria for selection.

ALS panelists' responses to process evaluation questions concerning the exemplar items are shown in Table 51. These were questions 19 and 20 on the last process evaluation questionnaire. Mean ratings were positive. Each was approximately 4.00 or higher on a scale of 1–5. This indicates a reasonably high level of satisfaction with the items selected as potential exemplars for illustrating performance at each level.

Table 51: Average ratings of responses to questions about exemplar items
(5 = *totally agree*, 3 = *somewhat agree*, 1 = *totally disagree*)

Round	Question #	Question	ALS Average Rating		
			Grade 4	Grade 8	Grade 12
Post	5-19	I believe the exemplar items will be useful for describing the achievement levels.	4.77	4.37	4.46
Post	5-20	The exemplar items I reviewed seemed appropriately matched to their achievement level.	4.41	3.89	3.96

ACT recommends that the Governing Board use the lists of items mapped to the achievement levels in the ALS meeting and the ALS panelist ratings of exemplars, along with other criteria of its choosing, to select exemplar items for the achievement levels.

SUMMARY AND CONCLUSIONS

For the purposes of helping the Governing Board set achievement levels for the 2009 NAEP in science for grades 4, 8, and 12, ACT:

- conducted a Pilot Study in which cut scores for the 2009 NAEP in science for grades 4, 8, and 12 were set using Mapmark with Whole Booklet Feedback;
- reviewed the results of the Pilot Study with the TACSS and identified ways to improve implementation of the methodology for the ALS meeting; and
- implemented the improved Mapmark with Whole Booklet Feedback methodology for the operational ALS meeting.

Data from the ALS meeting provided evidence of procedural validity, internal consistency, and reasonableness of the results. The ALS meeting received high ratings on the panelist process evaluation questionnaires across all categories including clarity of instructions, panelist understanding of tasks, and panelist understanding of the meaning of performance at the lower borderline of each achievement level. ALS panelists also indicated that they had sufficient time to complete their tasks. In addition, panelist ratings of the efficacy of the method in yielding reasonable cut scores were high and most panelists (83 out of 85) indicated they would sign a statement recommending the use of the resulting cut scores. These results indicate that the quality of the ALS procedure used for science was comparable to the quality of processes used to establish achievement levels for other NAEP subject areas.

The ALDs were also well received by the panelists. Panelist ratings of their understanding of the ALDs were high and increased across rounds. By the final round, they felt that their cut scores were highly consistent with the level of performance described in the ALDs.

The internal consistency measures used showed that the cut scores can be considered as reliable. There were no significant differences between mean cut scores by panelist type, race/ethnicity, gender, and geographic region, except between regions for the Proficient and Advanced cut scores for the final round for grade 8. The mean group cut scores were significantly different for grade 4 Proficient level in round 1, for grade 8 Advanced level in the final round, for grade 12 Basic level in round 1 and round 3, and for grade 12 Advanced level in round 1. Many of the table groups also showed significant differences. In round 3 this can be attributed primarily to small within-table-group variability. In round 1, both grade 4 and grade 12 had table level effects that were significant at two of the three achievement levels. Most of this can be attributed to the small sample sizes of the table groups and the effect of outliers on the mean.

ACT's TACSS reviewed the ALS meeting process and results, and concluded as summarized above that the procedural validity was strong, the ALS cut scores were reliable, and that the panelists' reactions to the consequences data provide support for the achievement levels. Based on these results, ACT recommends the cut scores from round 3 of the ALS meeting. The cut scores, on the ACT NAEP-like scale used at the ALS meeting, are in Table 52.

Table 52: Cut scores* for 2009 NAEP Science by grade and achievement level

Grade	Basic	Proficient	Advanced
4	328	376	447
8	570	598	647
12	785	820	866

*The ACT NAEP-like scales have means (SDs) of 364 (33), 579 (33), and 793 (33) for grades 4, 8, and 12, respectively.

ACT also recommends that the Governing Board use the lists of items shown for each achievement level in Appendix D along with panelists' ratings of these items as exemplars, plus other information such as item content and difficulty, in selecting exemplar items for NAEP reports. It is further recommended that the Governing Board consider most strongly those items that were rated by 50% of the panelists as *Very Good* and by fewer than 30% of the panelists as *Do Not Use*.

Based on these activities, ACT provided the Governing Board at their May 14, 2010, Board meeting with the following input regarding the three recognized outcomes of the Achievement Level Setting process:

- ACT endorses the ALDs that were used in the operational ALS meeting.
- ACT recommends the cut scores from round 3 of the operational ALS meeting. These cut scores would be transformed to the scales that will be used to report the 2009 science assessment results for grades 4, 8, and 12.
- ACT recommends that the Governing Board use the lists of potential exemplar items from the ALS meeting in the process of selecting exemplar items. Ratings of these items by ALS panelists should be taken into consideration in selecting exemplar items.

These recommendations and endorsements are based on positive evaluations and conclusions concerning relevant elements of the process by panelists and ACT's Technical Advisory Committee on Standard Setting.

Recommendations for Future Standard Settings

ACT has several recommendations for future standard-setting meetings. Although evaluations of the Mapmark with Whole Booklet Feedback method were overwhelmingly positive, there were some areas about which either panelists or ACT staff expressed some concern. ACT recommends the following changes to the process.

ACT suggests that the Board change the proportions of teachers, nonteacher educators, and general public to be included on a standard-setting panel. Current Governing Board policy is to have 55% teachers, 15% nonteacher educators, and 30% general public on the standard setting panels. While there are compelling reasons for including the general public, they are much more difficult to recruit. For the 2009 Science standard setting, nominations were received (across all grades) for 326 teachers, 244 nonteacher educators, and 104 general public. If we eliminate those who were rated in the lowest category of qualifications, that left 223 teachers, 181 nonteacher educators, and 63 general public. The problem is then that we are choosing almost all of the general public nominees and leaving out many excellent candidates from the nonteacher educator group. This latter group can include many of the best teachers who have been promoted to school, district, and state positions. We suggest a 55-30-15 or 60-20-20 teacher/nonteacher educator/general public split.

In the design of this study, part of the round 2 feedback consists of a selection of student booklets shown to the panelists. These booklets are selected at the round 1 median cut scores and in the middle of each achievement level (see Figure 18). The rationale behind selecting booklets at these points is to show the panelists what a minimally-qualified student can do for the specific achievement level, and to compare that with students scoring at a level that would indicate "solid" performance within an achievement level. Selecting booklets in this way seems reasonable, but carries with it a significant disadvantage. In particular, the booklets can only be selected after the cut scores are known, at the end of round 1. The selection and copying process is so lengthy that it requires that round 1 be completed at the end of a day. This severely constrains any flexibility in the schedule. To be more efficient and make it possible to end round 1 during the day, we suggest that the student booklets be selected and copies made prior to the ALS meeting. This could be done by selecting booklets at a fixed interval along the scale (e.g., every 20 points). During the meeting, provide panelists with copies of those booklets that range from the middle of Below Basic to the middle of Advanced. At a point or two along the scale (e.g., the Proficient cut score), select two booklets so that panelists will be able to evaluate the similarities and differences in performance of students scoring at the same level. The selection of the range for the booklets and the interval between the booklets must be made judiciously, to ensure that there are booklets at each of the achievement levels.

In round 3, the current design has a discussion of the results of the round 2 cut scores, followed by presentation of the consequences data giving student performance with respect to the round 2 cut scores (see page 43). At that point, there is a grade-group discussion of the results. This has not led to the robust discussion of results that was hoped for, and this phenomenon has been consistent across grade groups. This may be due to the fact that this comes near the end of the process when the panelists are somewhat fatigued, but, given the relevance of the consequences data to the Board's evaluation of the cut scores, it would be helpful to have more comments from the panelists about the appropriateness, or lack thereof, of the percentages of students at or above each of the achievement levels and to have a discussion of cut scores—why do you think the group cut score is just right, too low, or too high? It is possible that including a table-group discussion, as well as a grade-group discussion, of the round 2 consequences data would elicit more response.

In the Statement of Work for the ALS process, proposers were invited to suggest paying the pilot study and ALS panelists as part of the standard-setting process. ACT chose not to include payment to panelists as part of its proposal, as at the time we understood the proposal to say that this would be considered only if the proposer could provide documented evidence that paying panelists would improve the participation rate. (We now understand that this may have been a misreading of what was intended.) We were unable to find research documenting the effectiveness of paying panelists, and so felt unable to comply with this requirement. In any case, ACT was able to obtain nominations for a sufficient number of general public panelists without offering payment. However, we feel that science may be a unique case, and that in other subject areas it may again prove more difficult to recruit general public members of the panel. Thus, we feel that paying the panelists should continue to be an option made available to the contractor.

It has become more difficult to recruit panelists over time, and this is directly tied to the difficulty in soliciting nominations. The current method emphasizes sending out a large volume of requests for nominations, with the expectation that even if only a small percentage of those contacted respond with nominations, this will yield a sufficient number of nominees. If the response rate continues to decline, there will be a point beyond which it is impossible to get a high-quality representative panel. There also may be a perception that the NAEP standard setting process is regarded as generally unimportant, given that so few people bother to nominate anyone for the panels. We feel that, at least to this point, this is not the case. The science panelists seemed to be of high quality, and there has been no discontent expressed concerning the selection or make-up of the panels.

It may be possible to increase the response rate for the nominators by more carefully targeting the people that are contacted and trying to increase the importance of nominating someone to the person contacted. Some ideas for accomplishing this are:

- Work with the state NAEP coordinators to get nominations. These coordinators are already familiar with NAEP and should be highly motivated to assist. They should also have access to names and contact information for both highly qualified teachers, and nonteacher educators.
- Get a letter from the chief state school officer in the state, endorsing participation in the standard setting, or encouraging principals to nominate teachers. This letter could precede the letter asking for nominations or accompany it. It has been our experience that these types of letters typically increase the nomination and participation rates. However, the difficulty of getting the cooperation of these officials would have to be taken into consideration.
- Work with organizations that specialize in the subject area to get nominations. We already do this to some extent, but by working at a more individual level with the leaders of these organizations, it may be possible to work down to people at a more local level to get more nominations. This does run the risk of letting these organizations have too much influence on the achievement level results.

These ideas would be most likely to increase the number of nominations for the teacher and nonteacher educator groups, which are not the real problem. If the suggestion of changing the percentages were adopted, many of the recruiting problems would go away.

References

ACT, Inc. (2005). *Developing achievement levels on the 2005 National Assessment of Educational Progress in grade 12 mathematics: Process report*. Iowa City, IA: Author.

ACT, Inc. (2007). *Developing achievement levels on the 2006 National Assessment of Educational Progress in grade 12 economics: Process report*. Iowa City, IA: Author.

ACT, Inc. (2010). *Developing achievement levels on the 2009 National Assessment of Educational Progress in science for grades four, eight, and twelve: Technical report*. Iowa City, IA: Author.

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician, 37,* 36–48.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). Standard Setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard-setting procedures using behavioral anchoring.* Symposium conducted at the Council of Chief State School Officers National Conference on Large Scale Assessment, Phoenix, AZ.

Loomis, S. C. & Hanick, P. L. (2000). *Developing achievement levels for the 1998 NAEP in civics: Final report.* Iowa City, IA: ACT.

Maritz, J. S. and Jarrett, R. G. (1978). A note on estimating the variance of the sample median. *Journal of the American Statistical Association, 73*, 194–196.

Masters, G. N., Adams, R., & Lokan, J. (1994). Mapping student achievement. *International Journal of Educational Research*, *21*, 595–609.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: psychological perspectives. In G. J. Cizek (Ed.), *Setting Performance Standards*. Mahwah, NJ: Lawrence Erlbaum Associates.

National Assessment Governing Board (2008a). *Work Statement for Developing Achievement Levels on the 2009 National Assessment of Educational Progress for Science for Grades 4, 8, and 12.* (Attachment A: Statement of Work, Solicitation: ED-08-R-0028.) Washington, DC: Author.

National Assessment Governing Board (2008b). *Achievement Levels Policy and Implementation Guidelines.* (Appendix A in Attachment A: Statement of Work, Solicitation: ED-08-R-0028.) Washington, DC: Author.

National Assessment Governing Board (2008c). *Science Framework for the 2009 National Assessment of Educational Progress.* Washington, DC: Author.